CapAtlas: Resolution-Specific Validity of Sparse Behavioral Model Maps

Anonymous authors
Paper under double-blind review

Abstract

Thousands of language-model checkpoints are poorly documented, yet comprehensively benchmarking every model is impractical. Sparse model maps offer a cheaper alternative, but current evaluations collapse map quality into one score and leave their downstream validity unclear. We introduce *CapAtlas*, a leakage-controlled protocol that learns a capability map from documented references, freezes it, places an unseen model from 16 response distributions, and only then unseals its evaluation profile. We evaluate a 46-model atlas on 70 targets from 43 unseen lineages, using 84,000 held-out response records and 109,480 sealed oracle outcomes for evaluation. Reference supervision improves graded local retrieval from 0.8181 to 0.8609 lineage-macro NDCG@5 (+0.0428; 95% CI $[0.0229, 0.0640]$; random floor 0.6048), with a positive effect from checkpoint- to architecture-level aggregation. Six calibration summaries reach 0.8611, showing that calibration explains the aggregate score while itemwise geometry contributes selectively by domain. Under a disjoint capability construct, global ranking improves by 0.1691 Kendall's τ; the registered local-retrieval gate identifies a separate neighborhood-calibration target. On 50 prospective targets, a frozen reliability score stratifies placement risk (AUROC 0.731, 95% CI $[0.582, 0.866]$). The operating point reaches 80.95% precision; holding its accepted set fixed, two oracle halves yield 80.95% and 85.71%, both below the registered 90% standard and defining a reproducible calibration target. CapAtlas turns sparse model mapping into an auditable instrument whose claims are matched to their decision resolution.

1 Introduction

A practitioner who downloads an undocumented checkpoint faces a cold-start triage problem: its card may be missing, inherited from a sibling variant, or reported on benchmarks incomparable with anything else in the registry (Horwitz et al., 2025), yet they must still decide which known models it resembles and which expensive evaluations to pay for next. Exhaustive benchmarking defeats the purpose, so a scalable capability layer must work from a small amount of fresh evidence.

Asking an unknown model a short standardized set of questions recovers real structure: responses cluster models by family, vintage, and skill. That is not enough for the decision at hand. A map can place a model roughly right yet return the wrong nearest neighbors, and a score that flags shaky placements can order easy and hard cases correctly yet still be too error-prone to authorize an accept/reject rule. One average similarity number cannot separate these cases.

The practical question is therefore not whether a map is accurate but whether it is accurate enough for the decision made from it. We study four jobs one placement might do: order the registry coarsely, return a short list of nearest peers, flag which of its own placements are risky, and accept at a precision a user could rely on. Each job needs its own evidence, so a map earns four verdicts rather than one label (Figure 1).

CapAtlas is the protocol in that figure. Documented references are the only source of supervision; every learned object—panel, standardizer, projection, reliability head, threshold—is fixed before any target is seen; and a target from an unseen lineage supplies nothing but the same 16 probe distributions, with its ranking, peers, and risk score written down before its evaluation profile is opened.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

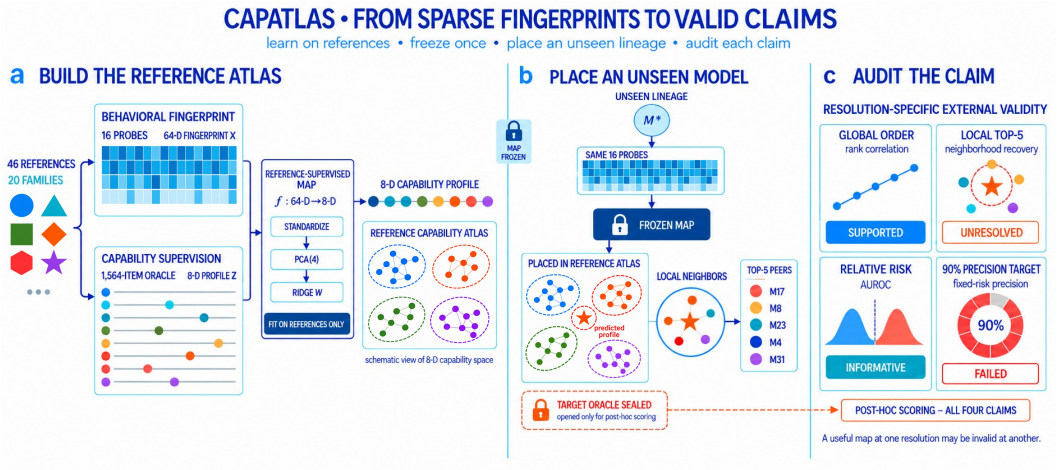


Figure 1: CapAtlas builds a reference-supervised behavioral map and uses it to place an unseen model. References provide sparse response distributions and independently measured capability profiles; a target supplies only the same 16 responses, its oracle sealed until scoring. Panel (c) carries one verdict per decision resolution. Its UNRESOLVED badge refers to *exact* Top-5 membership under a changed capability definition (Section 5.4); the *graded* local endpoint, NDCG@5, improves under the primary definition (Section 5.1).

The central object of study is the map's validated claim set. A transparent PCA–ridge chart serves as the controlled instrument, while alternative supervised geometries and aggregate controls distinguish general mapping evidence from estimator-specific behavior.

The resulting validity profile forms a clear resolution ladder. Reference supervision improves graded local peer retrieval on unseen lineages and across estimator families, while aggregate calibration explains much of that overall score and itemwise geometry contributes selectively by domain. Global ordering transfers to an independent capability construct, and risk ranking remains informative on a prospective cohort. Exact Top-5 membership and 90% fixed-precision acceptance become explicit calibration targets. We contribute:

- *unseen-lineage sparse placement* as an estimation problem with an explicit estimand, making model-map claims falsifiable;
- a *leakage-controlled protocol* that keeps target evaluation results out of map construction, so a placement is third-party auditable;
- a *resolution-specific validity framework* with formal analyses that explain why useful ranking need not imply reliable neighborhood recovery or fixed-precision certification;
- *prespecified validation* under lineage shift, a changed capability definition, and a prospective cohort, delivering independently testable validity claims at each decision resolution.

2 Related Work

Model maps and behavioral similarity. The Model Atlas proposal represents models through attributes, relations, and transformations (Horwitz et al., 2025). Behavioral distances recover model relationships and predict performance (Yax et al., 2025); few-query fingerprints identify model versions across vendors (Pasquini et al., 2025); shared errors reveal rising behavioral correlation (Goel et al., 2025); and label-free maps from 10,000 shared prompts recover family structure and temporal trends (Qiao et al., 2026). These studies establish that behavior carries map structure and score it globally; CapAtlas instead asks which downstream claims that structure licenses under sparse observation and atlas-disjoint lineages.

Sparse evaluation and capability prediction. tinyBenchmarks selects small evaluation subsets with item-response models (Maia Polo et al., 2024); metabench, General Scales, and BenchPress learn latent or predictive structure from model–item result matrices (Kipnis et al., 2025; Zhou et al., 2026; Zeng & Papailiopoulos, 2026); Anchor Points predicts aggregate rankings and instance outcomes

from few examples (Vivek et al., 2024); and other work selects evaluations, learns acquisition policies, or condenses samples (Zhou et al., 2025; Li et al., 2025; Rubinstein et al., 2026; Zhang et al., 2026). DKPS is the closest comparator, embedding cached target and reference responses and predicting scores by regression or nearest neighbors (Helm et al., 2026); with PLS and metric learning these are adjacent estimators of the same structure. CapAtlas shifts the objective from picking one regressor to learning and validating a reference-supervised capability geometry: peer placement is the estimand, its oracle is disjoint from the probes, every target lineage is excluded from the atlas, and validity is reported per decision resolution.

Low-dimensional capability structure and selective prediction. Scaling laws and factor analyses suggest that capability has low-dimensional structure (Ruan et al., 2024; Burnell et al., 2023), motivating the low-rank maps tested here, while compact tests and learned evaluators motivate direct measurement of fine-ranking fidelity (Yauney et al., 2026; Zhang et al., 2025). Connecting such a geometry to a deployment decision needs a reliability score and a threshold, whose ranking quality and fixed-precision behavior have distinct calibration requirements.

Appendix A summarizes these distinctions across the closest lines of work.

3 The CapAtlas Protocol

CapAtlas has two stages (Figure 1): Stage I builds a reference-supervised map and freezes it; Stage II places an unseen target through the same sparse interface, recording its outputs before any capability outcome is seen.

3.1 Task and Claim Resolutions

Let the reference atlas be $\mathcal{A} = \{(\mathbf{p}_i, \mathbf{z}_i, f_i)\}_{i=1}^{N}$ with $\mathbf{p}_i = (\mathbf{p}_{iq})_{q=1}^{J}$, where $\mathbf{p}_{iq} \in \Delta^{C_q-1}$ is model i's candidate-probability vector on probe q ($C_q = 4$ in our four-choice interface), $\mathbf{z}_i \in \mathbb{R}^D$ ($D = 8$) an independently measured capability profile under a declared capability definition, or *ontology*, O, and f_i a declared lineage. For target m we observe only the fingerprint $\phi_S(m)$, concatenated over a fixed panel $S \subset [J]$ with $|S| = B$, while $\mathbf{z}_m$ stays sealed until scoring. Profiles are standardized coordinate-wise within the eligible references, $\tilde{\mathbf{z}} = T_{\mathcal{A}}(\mathbf{z})$, and $d_z(i,j) = \|\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j\|_2$ is the oracle capability distance. The estimand is the ordering of references by $d_z(m,\cdot)$, which the sealed profile withholds at scoring time. The primary regime is lineage-disjoint: supporting analyses restrict to $I_{-f_m} = \{i \in [N] : f_i \neq f_m\}$, and in the evaluated cohorts the exclusion is automatic. The four jobs of Section 1—global order, local neighborhood recovery, relative placement risk, and acceptance at a threshold fixed in advance—are four functionals of this one placement. We score neighborhood recovery both graded, by NDCG@5, and exactly, by Recall@5, and never read one as the other.

3.2 Learning a Reference-Supervised Capability Geometry

The framework estimates a conditional-mean capability map that several supervised estimators can instantiate. For ontology O and reference population $P_{\mathcal{A}}$, let

$$m_O(x) = \mathbb{E}_{P_{\mathcal{A}}}[\tilde{\mathbf{z}} \mid \phi_S = x], \qquad d_{m,O}(x,x') = \|m_O(x) - m_O(x')\|_2. \tag{1}$$

This is a pseudometric: fingerprints with the same conditional-mean capability are equivalent even when their raw responses differ, and the estimand above is the ordering it induces. Stage I instantiates m_O with a plain plug-in, using references only: select the probe panel by a reference-fitted item-response criterion excluding oracle coordinates; standardize fingerprints and profiles; fit PCA and retain four components $\mathbf{u}_i = P_4(\phi_S(i))$ without rescaling; and fit a multi-output ridge decoder

$$(\hat{W}, \hat{\mathbf{b}}) = \arg\min_{W,\mathbf{b}} \sum_{i \in I_*} \|\tilde{\mathbf{z}}_i - W\mathbf{u}_i - \mathbf{b}\|_2^2 + \lambda \|W\|_F^2, \qquad \lambda = 10, \tag{2}$$

with $I_* \subseteq [N]$ the eligible reference set and $I_* = [N]$ in the reported evaluation (ϕ_S has 64 dimensions at $B = 16$). We call the rank-four plug-in $\hat{m}_O = \hat{W}P_4 + \hat{\mathbf{b}}$ SoftMap; neither $\lambda = 10$ nor the rank was tuned on any target (Appendix D). PLS, CCA, and full ridge estimate m_O under different assumptions, while metric-learning and DKPS variants target the induced ordering directly; agreement across them tests the capability geometry rather than one estimator choice.

3.3 Placement and Leakage-Controlled Audit

Stage II fixes what a new model reveals and when (Algorithm 1). The target supplies the same 16 probe distributions and nothing else; the fixed map predicts $\hat{\mathbf{z}}_m$, ranks references by $\hat{d}(m, j) = \|\hat{\mathbf{z}}_m - \tilde{\mathbf{z}}_j\|_2$ into an ordering and a Top-K peer set, scores the placement with a reliability head $r = g(x^*, \mathcal{A})$ built from target–atlas geometry alone, and accepts at a threshold τ set in advance (Section 4.4). Only after every output is written down is $\mathbf{z}_m$ unsealed.

Algorithm 1 Leakage-controlled CapAtlas fitting, placement, and audit

Require: reference fingerprints $X_{\mathcal{A}}$, reference profiles $Z_{\mathcal{A}}$, panel S, target responses X_*, peer count K, selector (g, τ)

1: Fit standardizer $T_{\mathcal{A}}$ and PCA P_4 on references only; fit ridge $(\hat{W}, \hat{\mathbf{b}})$ from $P_4(X_{\mathcal{A}})$ to $T_{\mathcal{A}}(Z_{\mathcal{A}})$; then **freeze** $S, T_{\mathcal{A}}, P_4, (\hat{W}, \hat{\mathbf{b}}), g, \tau$.
2: Observe only $\phi_S(X_*)$, the target's 16 response distributions; compute $\hat{\mathbf{z}}_* = \hat{W} P_4(\phi_S(X_*)) + \hat{\mathbf{b}}$.
3: Rank eligible references by $\|\hat{\mathbf{z}}_* - T_{\mathcal{A}}(\mathbf{z}_j)\|_2$, return Top-$K$ peers, compute $r_* = g(X_*, \mathcal{A})$ from atlas-only stability features, and accept iff $r_* \geq \tau$.
4: Save the profile, ranking, peers, score, and decision before accessing the target oracle.
Post-hoc scoring only:
5: Unseal $\mathbf{z}_*$ and evaluate global order, Top-K recovery, risk ranking, and fixed precision.

The primary comparison holds the target's information fixed and changes only how the geometry is built: the matched-information control ranks reference j by the mean itemwise Jensen–Shannon distance $d_{\mathrm{JS}}(m, j) = |S|^{-1} \sum_{q \in S} \sqrt{\mathrm{JS}(\mathbf{p}_{mq}, \mathbf{p}_{jq})}$ with natural logarithms. Soft-JS keeps the complete response distributions and gives an estimation-free geometry in response space; SoftMap learns the corresponding geometry in capability space. Because both read the same 16 distributions, a difference comes from how the geometry is built, not from what the target discloses; Section 5.3 decomposes that construction only partly. Placement costs 16 target queries against a one-time atlas oracle.

3.4 Why Validity Separates across Resolutions

Let $T_K(t)$ be the oracle Top-K references for target t, $d_{(K)}$ its Kth oracle distance, and $\delta_t = \|\hat{\mathbf{z}}_t - \tilde{\mathbf{z}}_t\|_2$ the prediction error.

Proposition 1 (Prediction-error-controlled Top-K recovery)**.** *If every $j \notin T_K(t)$ satisfies*

$$d_z(t, j) - d_{(K)} > 2\delta_t, \tag{3}$$

then ranking references by $\hat{d}(t, j)$ recovers $T_K(t)$ exactly.

Exact Top-K recovery therefore needs the oracle boundary margin to exceed twice the prediction error, a requirement coarse ordering does not carry; this is why global order can improve while local peers stay unstable.

A second separation applies to the decision-level claims. Let $Y = 1$ denote a resolved placement, R a reliability score, a its AUROC, $\theta = \Pr(Y = 1) \in (0, 1)$, and a fixed threshold have coverage $\kappa > 0$ and selective precision γ.

Proposition 2 (Rank discrimination and precision require separate guarantees)**.** *Every such operating point obeys*

$$\frac{\kappa(1 - \gamma)}{1 - \theta} \left(1 - \frac{\gamma\kappa}{\theta}\right) \leq 1 - a. \tag{4}$$

AUROC is invariant to strictly increasing score transformations and to pure class-prior shift that leaves the class-conditional score laws fixed, whereas precision at a fixed threshold is not. Hence rank discrimination and absolute precision describe complementary properties of the score.

The envelope constrains precision without determining it, so a score can rank failures well and still miss a precision target. Proofs, a graded Recall@K bound, and a decomposition of δ_t into discarded capability, nuisance, and response noise are in Appendix S. Neither proposition bounds effect sizes, but together they predict that broad order tolerates perturbation better than local membership: without ties, one pair inversion moves Kendall's τ by $2/\binom{N}{2}$ while one Top-K boundary crossing moves Recall@K by $1/K$. We test that prediction under three capability definitions and two cohorts.

4 Experimental Design

4.1 Models, Probe Interface, and Capability Ontologies

The atlas has 46 checkpoints from 20 declared families; fitting, panel construction, standardization, and every selection decision use only these references, and a timestamped protocol fixed targets, taxonomy, panels, oracles, methods, contrasts, and decisions before evaluation. The primary cohort is 70 of 72 smoke-valid checkpoints over 43 declared lineages. Each completed 1,200 response rows (240 candidate-panel items plus 960 disjoint audit items) and 1,564 sealed oracle items, so the 70 targets alone account for all 84,000 response rows and all 109,480 oracle-item rows. That volume buys evaluation coverage, not placement cost: each placement still reads 16 response distributions, and inference remains lineage-macro. The prospective cohort adds 50 checkpoints from 45 lineages. **Every target lineage in both cohorts is absent from the 46-model atlas, and the 50 prospective targets share neither a checkpoint nor a declared lineage with the 70 primary targets used to fit the reliability selector.** A declared lineage is an author-assigned ancestry group fixed in the manifest before scoring: derived checkpoints inherit their base-model group, while separately recorded pretraining generations stay separate (Table 9). The grouping is operational rather than canonical, and we macro-average by it throughout; identifiers, hashes, attrition, and execution history are in Appendices C–F. The probe interface is fixed: each model answers 240 candidate and 960 disjoint MMLU audit items across four domains (Hendrycks et al., 2021a), a reference-only item-response model selects four probes per domain for the static 16-probe panel, and SoftMap and soft-JS receive identical four-way probabilities (Appendix E).

Three ontologies instantiate the protocol: the primary O_1 is an eight-coordinate profile over 1,564 items per model from PubMedQA, GSM8K, MATH, MBPP, HumanEval, CaseHOLD, ARC-Challenge, and HellaSwag (Jin et al., 2019; Cobbe et al., 2021; Hendrycks et al., 2021b; Austin et al., 2021; Chen et al., 2021; Zheng et al., 2021; Clark et al., 2018; Zellers et al., 2019). The benchmark-disjoint O_2 replaces it with six tasks sharing no benchmark with O_1 (Lin et al., 2022; Sakaguchi et al., 2020; Bisk et al., 2020; Talmor et al., 2019; Clark et al., 2019; Welbl et al., 2017), on the same atlas, targets, and probe responses; the prospective reliability ontology O_3 uses QASC, RACE, SWAG, and SuperGLUE's COPA, RTE, and WSC (Khot et al., 2020; Lai et al., 2017; Zellers et al., 2018; Wang et al., 2019), sealed independently of the cohort that fit the selector. Under O_2 and O_3 the decoder is refit on the 46 reference profiles only—*reference-only refitting under an independent capability construct*, not transfer of a fitted decoder. Target oracle values stay hidden until rankings are complete in all three cases.

4.2 Comparators and Mechanism Controls

Comparators are organized by the alternative explanation each tests; the response-space baseline is soft-JS (Section 3.3). *Scalar-strength* controls—16-item accuracy, and retrieval after removing oracle PC1—ask whether the geometry reduces to one number. *Aggregate* controls—calibration-only, nuisance-only, metadata-only, and parameter count—ask whether itemwise content is needed at all. The *supervised estimator family*—hard and full-soft ridge, PLS, CCA, metric learning, BenchPress- and PhyloLM-style predictors, and DKPS regression and neighbor variants (Helm et al., 2026)—tests whether one algorithm is responsible. *Local-identification* controls restrict retrieval itself, to 15 ability-matched references or to single domains. Full specifications, residualization, scaling analyses, and a post-hoc decision-cost analysis are in Appendices D and M.

4.3 Endpoints, Statistical Inference, and Decision Gates

The primary endpoint is lineage-macro NDCG@5; Recall@5 (chance 0.109), Kendall's τ, top-1 regret, and capability RMSE are secondary. Oracle relevance is $\mathrm{rel}(m, j) = \exp\{-d_z(m, j)/s_m\}$ with s_m the median target-to-reference distance, and DCG@K is normalized by the oracle-distance ordering, so every reference has positive relevance and NDCG@5 carries a high chance level. It must therefore be read against explicit references: a uniform-random floor, an atlas-centroid floor, and a split-half ceiling that ranks by one oracle half and scores against the other (100 items per half; 82 for HumanEval). Each half is noisier than the full oracle, so this ceiling indexes oracle reliability rather than bounding attainable performance (Appendix O).

The three prespecified paired contrasts use one million lineage sign flips, Holm correction across the contrast family, and a 60% lineage sign-majority requirement. Uncertainty uses two resampling units we never conflate: the *lineage bootstrap* resamples declared lineages, and the *nested item-by-lineage bootstrap* resamples oracle items within task and then lineages, giving two separately reported estimates of the second-ontology Kendall effect (+0.1691 and +0.1731). Unit tests enforce scoring symmetry, item alignment, reference-only fitting, and lineage exclusion.

Under O_2, the frozen co-primary local-retrieval rule requires gains over soft-JS and calibration-only, while global Kendall receives its own secondary resolution-specific verdict.

For the decision-level claims a target is *locally resolved* when at least three of the five retrieved references are in its oracle Top-5 (Recall@5 ≥ 0.6); the prospective certificate requires every registered condition in Table 3 at once, so it passes or fails as a whole.

4.4 Reliability Head and Operating Point

The head g of Section 3.3 uses six target–atlas stability features computed without any oracle value: normalized Top-1 and Top-5 boundary gaps, normalized nearest-atlas distance, leave-one-reference-out Top-1 consensus and Top-5 Jaccard, and leave-one-reference-out prediction spread. A class-balanced logistic regression ($C = 1$) and its standardizer are fit on completed primary-cohort targets using the O_2 local-resolution outcomes only; τ is the largest development acceptance prefix with at least five targets and empirical precision at least 80%, and all three are serialized before any prospective target is scored. The prospective success criterion is a separately registered 90% standard, stricter than the 80% rule that placed τ: the gate asks whether the operating point clears 90%, not whether calibration met its own target.

5 Results: A Resolution-Specific Validity Profile

The five findings map onto the four claim resolutions of Section 3.1. Sections 5.1–5.3 test *local peers* on the primary cohort under O_1—whether the gain exists, whether it depends on one estimator, and what carries it. Section 5.4 refits on references under O_2 and separates *global order* from exact local membership; Section 5.5 tests *relative risk* and *fixed-precision acceptance* prospectively under O_3.

5.1 Reference Supervision Improves Graded Local Retrieval on New Lineages

Verdict: the primary claim is supported. Under O_1, learning the geometry from reference profiles retrieves better peers than matching raw responses, on lineages the atlas has never seen.

Across 70 targets from 43 atlas-disjoint lineages, SoftMap reaches 0.8609 lineage-macro NDCG@5 against 0.8181 for soft-JS. The paired improvement is 0.0428 (descriptive lineage-bootstrap 95% interval $[0.0229, 0.0640]$; two-sided sign-flip $p = 1.25 \times 10^{-4}$, Holm 3.75×10^{-4}); 31 of 43 lineages improve, clearing the prespecified 60% sign majority, and Figure 2 shows a broad lineage-level shift rather than a few large gains. Exact membership moves the same way—retrospective Recall@5 0.4271 against 0.3853—but that contrast has no prespecified interval and is not a tested endpoint (Appendix G). Random ranking scores 0.6048 and atlas-mean ranking 0.6873, so soft-JS already carries real capability information; against the 0.9556 split-half ceiling the map closes 0.043 of the remaining 0.137, or 31% of the measurable gap at the local-peer resolution.

The primary effect is stable across the frozen taxonomy. Reaggregating the unchanged target scores from 70 checkpoints to 43 declared lineages, 41 publisher namespaces, and 20 implementation classes leaves a positive interval and the sign-majority criterion at all four resolutions, with mean gains of 0.0356 to 0.0429; the most severe coarsening gives 0.0356 ($[0.0098, 0.0606]$, exact sign-flip $p = 0.0166$). Clustering dependence by architecture instead gives $[0.0169, 0.0554]$ ($p = 0.00835$), and the worst single-lineage deletion and all 260 admissible one-pair merges keep positive lower bounds (Appendix H).

5.2 Supervised Capability Geometries Consistently Improve Retrieval

Verdict: the retrieval gain is stable across supervised geometry estimators. PLS, metric learning, DKPS, full soft ridge, and SoftMap all sit above soft-JS in one narrow band (Table 1), and no

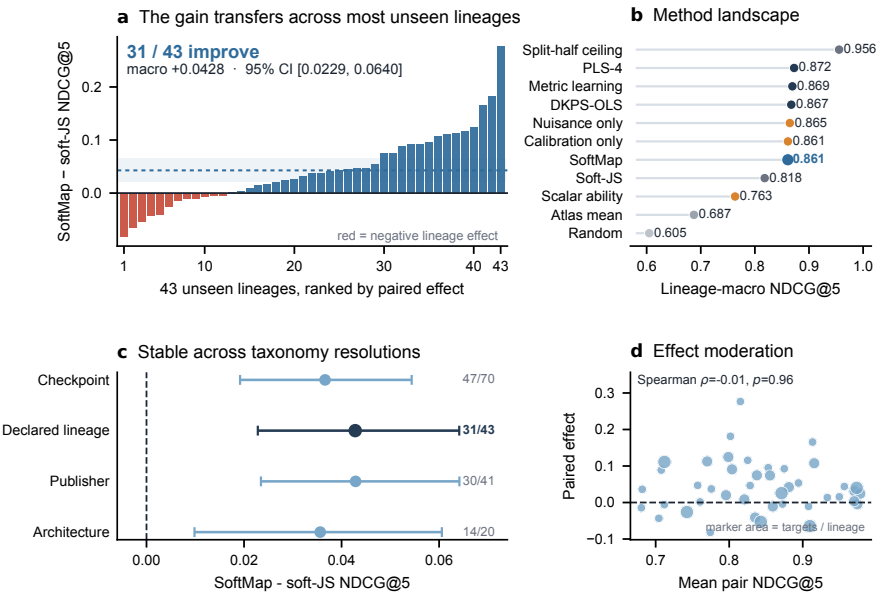


Figure 2: Reference supervision improves graded local retrieval in 31 of 43 atlas-disjoint lineages. (a) Lineage effects and descriptive 95% interval. (b) Methods against two floors and split-half oracle reliability. (c) The unchanged target scores retain a positive effect across four outcome-independent taxonomy resolutions; horizontal bars are descriptive 95% group-bootstrap intervals. (d) Effect versus pair performance; marker area denotes target count.

Table 1: Selected lineage-macro NDCG@5 over 43 unseen lineages (70 targets; 46-model atlas), descending down the left column then the right. SoftMap is prespecified; post-freeze comparators are descriptive, the aggregate controls come from a separately hashed recomputation, and sub-thousandth differences do not resolve the descriptive estimator band (Appendix G).

Method	NDCG@5 ↑	Method	NDCG@5 ↑
Split-half oracle ceiling	.9556	SoftMap (prespecified)	.8609
PLS, rank 4	.8725	Full soft ridge	.8586
Metric learning	.8693	Soft-JS	.8181
DKPS, OLS	.8672	Scalar ability	.7634
Nuisance only (retrospective)	.8646	Atlas-mean floor	.6873
Calibration only (retrospective)	.8611	Random floor	.6048

method-to-method contrast survives correction (Holm $p \geq 0.673$). The data therefore identify a stable supervised-geometry band, with PCA plus ridge as its transparent prespecified instance.

5.3 The Geometry Combines Global Calibration with Domain-Specific Itemwise Signal

Verdict: the geometry combines an aggregate calibration signal with domain-specific itemwise structure. The controls identify the resolution at which each component contributes.

Two controls remove scalar strength: removing oracle PC1 leaves a SoftMap-minus-soft-JS gain of 0.0893 (descriptive 95% interval $[0.0594, 0.1214]$), and restricting retrieval to 15 ability-matched references leaves 0.0408 ($[0.0157, 0.0666]$). Both intervals exclude zero, so the signal survives removal of the dominant strength axis. Aggregate controls then bound what the learned geometry adds. Six response summaries—mean entropy, mean maximum probability, and the four mean answer-option probabilities—with and without coarse configuration metadata, reach the same 0.861–0.865 band as SoftMap (Table 1, Figure 3a). At the aggregate local-retrieval resolution, calibration alone is therefore sufficient. Domain restriction then locates the incremental itemwise signal: it adds value in biomedicine and general reasoning, while law, mathematics, and code stay inside the aggregate-control band (Figure 3b). The contribution is the resolution at which each component becomes useful.

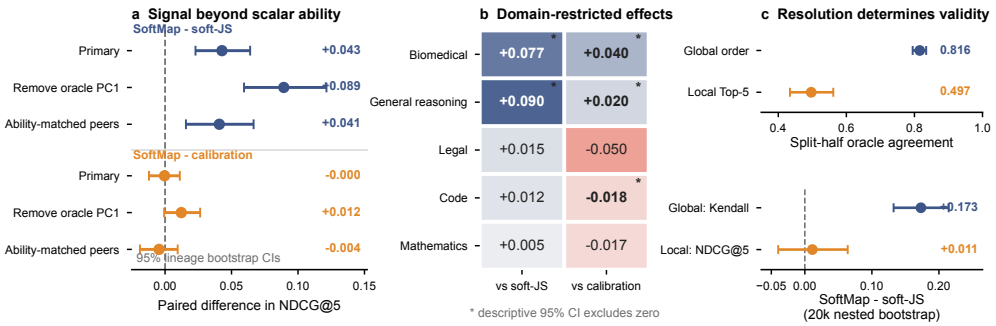


Figure 3: Mechanism controls locate the signal. (a) Retrospective paired lineage effects with descriptive 95% intervals, for SoftMap against soft-JS and against calibration-only. (b) Domain-restricted effects; asterisks mark intervals excluding zero. (c) The independent construct O_2 (Section 5.4): split-half oracle agreement is far higher for global order than for Top-5 membership, and a 20,000-draw nested bootstrap shows the same split in the estimated effects.

Table 2: Reference-only refitting under the six benchmark-disjoint tasks of O_2, same 70 targets and 43 lineages. Global order separates the methods; local estimates remain within uncertainty.

Method	NDCG@5 ↑	Recall@5 ↑	Kendall τ ↑
Calibration only	.8372	.4636	.7802
Full soft ridge	.8355	.4512	.7673
SoftMap	.8310	.4481	.7640
Soft-JS	.8230	.4473	.5949
Scalar ability	.6089	.0907	.4471

5.4 Global Ordering Transfers Across an Independent Capability Construct

Verdict: reference-only refitting transfers global ordering across a benchmark-disjoint capability definition and isolates local peer recovery as a separate calibration target.

The two resampling units agree on the pattern and must be read separately. Under the lineage bootstrap, global order improves over soft-JS by 0.1691 Kendall's τ ($[0.1335, 0.2061]$), whereas graded local NDCG@5 changes by only 0.0080 ($[-0.0423, 0.0607]$) and exact Recall@5 by 0.0008 ($[-0.0930, 0.0992]$), with 20 of 43 lineages positive and Holm $p = 1.0$. The frozen co-primary local rule therefore fails: the NDCG gain over soft-JS is unresolved, and calibration-only is statistically indistinguishable from SoftMap. Positive global ordering is a separate resolution-specific finding. Under the nested item-by-lineage bootstrap, which additionally resamples oracle items within task, Kendall's τ improves by 0.1731 ($[0.1324, 0.2145]$) while NDCG@5 changes by 0.0112 ($[-0.0398, 0.0639]$; Figure 3c). Table 2 reports method-level endpoints, on which calibration-only is descriptively highest.

The oracle itself marks the boundary: split-half agreement within this construct is high for distance order ($\tau = 0.8155$) but only moderate for Top-5 membership (Jaccard 0.4968, $[0.4343, 0.5615]$), so the instrument resolves coarse order far better than neighborhoods—the margin condition of Proposition 1 that global order need not satisfy. Ordering transfers; exact peer membership becomes a specified ontology-specific calibration target.

5.5 Prospective Risk Ranking Is Informative and Defines a Recalibration Target

Verdict: the frozen score stratifies prospective placement risk and turns fixed-precision operation into a measurable recalibration target.

All 50 prospective targets from 45 lineages complete, and the score fixed in advance provides prospective risk ordering: AUROC is .731 ($[.582, .866]$). Accepted targets resolve 80.95% of the time versus 41.4% for rejected ones (odds ratio 6.02; one-sided Fisher $p = .0054$; Figure 4a). The selector accepts 21 targets, 17 locally resolved: 80.95% precision at 42.0% coverage (95% lower bound 61.56%). This nearly reproduces development (80.65% at 44.29% coverage) but misses the registered 90% requirement by 9.05 points while the other five conditions pass (Table 3, Figure 4b). At

Table 3: Registered prospective gate (50 targets). The 90% fixed-precision criterion fails by 9.05 points; the other diagnostics locate where the score remains informative. The LCB is compared with the 58.0% event rate; lineage success uses lineage mean ≥ 0.5.

Diagnostic	Observed	Requirement	Result
Targets / lineages	50 / 45	≥ 30 / ≥ 20	Pass
Coverage	42.0%	≥ 25%	Pass
Precision	80.95%	≥ 90%	Fail
Precision 95% LCB	61.56%	> 58.0%	Pass
Lineage success	80.0%	≥ 80%	Pass
AUROC	.731	Diagnostic	—

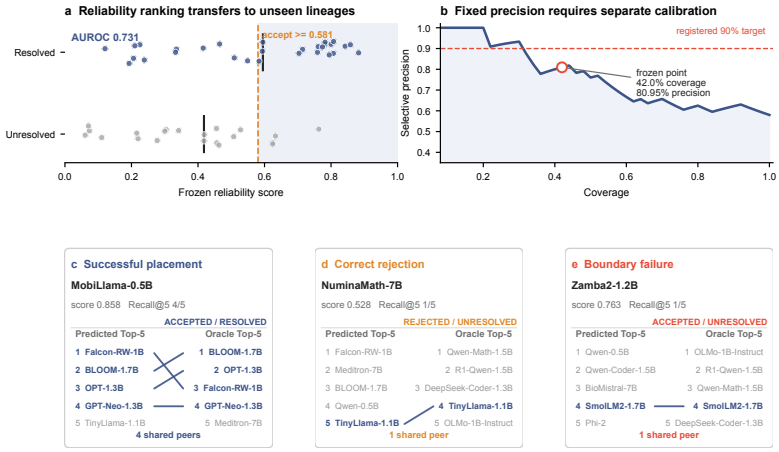


Figure 4: Prospective reliability. (a) Risk ranking is informative (AUROC .731). (b) The fixed operating point reaches 80.95% precision at 42.0% coverage, quantifying the gap to the registered 90%. (c–e) Representative outcomes; the complete 50-target ledger is Appendix K.

the observed AUROC, coverage, and event rate, Proposition 2 permits 55.2–100% selective precision, so the shortfall is a calibration target for the next prospective cohort.

A post-freeze oracle audit verifies that this target is stable: even–odd halves give distance-vector Kendall's $\tau = 0.8676$, Top-5 Recall 0.7489, and 91.11% agreement on the certificate label. Holding the frozen 21-target accepted set fixed, half-precision is 80.95% and 85.71%, below 90% in both splits. This localizes the remaining work to operating-point calibration (Appendix J).

6 Discussion

CapAtlas establishes a decision-indexed validity profile. Supervision improves graded local retrieval across unseen lineages and taxonomy resolutions; calibration recovers the aggregate band, while itemwise geometry contributes by domain. Global ordering transfers to an independent construct, and prospective scores stratify risk and quantify the recalibration required for fixed-precision operation. The prospective endpoint is stable across oracle halves, making 90% acceptance a concrete calibration target. Each claim should name its population, capability definition, interface, resolution, and uncertainty unit (Appendix B).

Scope and next validation. The evidence covers a 46-model atlas and a multiple-choice probability interface. Generative, agentic, long-context, and out-of-range targets require panel revalidation: reselection Jaccard is 0.393, and formatting and tuning affect retrieval (Appendix Figure 5). Larger open-set cohorts and prospective recalibration are the next steps toward fixed-precision operation.

Conclusion. CapAtlas makes sparse behavioral mapping leakage-controlled and decision-aware. Supervised maps improve graded peer retrieval, while calibration and domain analyses locate the signal. An independent construct and a prospective cohort support global ordering and risk ranking while quantifying the path to fixed-precision operation. Model maps should be judged at the resolution of their intended decision.

Reproducibility Statement

Code and derived artifacts are publicly available at `https://anonymous.4open.science/r/CapAtlas-96F1/`. The repository contains the reference implementation, sanitized cohort manifests, frozen protocols, checksums, per-target derived measurements, bootstrap summaries, and scripts that verify the released results and rerun the hierarchical-lineage analysis. Benchmark content and model outputs governed by third-party licenses are not redistributed. Stored seeds make panel construction and resampling deterministic. Post-freeze corrections retain before/after source hashes and leave the oracle, panels, methods, and success rule unchanged.

Ethics Statement

Capability mapping may assist documentation, but it can also profile systems without their operators' consent or overstate model suitability. A predicted location is neither a safety certificate nor a deployment recommendation. Fingerprints may encode contamination, cultural bias, or prompt-format artifacts; we therefore recommend reporting oracle and interface sensitivity and not using the present score for open-set rejection. No human-subject or private user data were used.

Data availability. The public repository provides sanitized manifests, probe identifiers, derived profiles, analysis artifacts, and aggregate results. Model weights, benchmark content, and restricted response rows remain subject to their source licenses; the released scripts and identifiers support reconstruction from original sources by licensed users.

Disclosure. Anonymous for review; CRediT, conflicts, and funding will be added camera-ready. AI-assisted tools supported literature retrieval, research ideation, code inspection, and manuscript drafting and polishing. The authors independently verified all claims, calculations, citations, and final text and take full responsibility for the work.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *International Conference on Machine Learning*, volume 267, pp. 19621–19678, 2025. URL `https://arxiv.org/abs/2502.04313`.

Hayden Helm, Ben Johnson, and Carey Priebe. Query-efficient model evaluation using cached responses. *arXiv preprint arXiv:2605.07096*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS 2021 Datasets and Benchmarks Track*, volume 1, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. We should chart an atlas of all the world's models. In *Advances in Neural Information Processing Systems, Position Paper Track*, 2025.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, 2019.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020. doi: 10.1609/aaai.v34i05.6319.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. metabench - a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, 2025.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017. doi: 10.18653/v1/D17-1082.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *International Conference on Machine Learning*, volume 267, pp. 35581–35602, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning*, volume 235, pp. 34303–34326, 2024.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *USENIX Security Symposium*, 2025.

Dong Qiao, Chris Ding, and Jicong Fan. Mapping and measuring the behavioral evolution of large language models. *arXiv preprint arXiv:2608.11027*, 2026.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.10938.

Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. DISCO: Diversifying sample condensation for efficient model evaluation. In *International Conference on Learning Representations*, 2026.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1576–1601, 2024.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations*, 2026.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. PhyloLM: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In *International Conference on Learning Representations*, 2025.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104, 2018. doi: 10.18653/v1/D18-1009.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Yuchen Zeng and Dimitris Papailiopoulos. You don't need to run every eval. *arXiv preprint arXiv:2606.24020*, 2026. URL https://arxiv.org/abs/2606.24020.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In *Advances in Neural Information Processing Systems*, 2025.

Taolin Zhang, Hang Guo, Wang Lu, Tao Dai, Shu-Tao Xia, and Jindong Wang. SparseEval: Efficient evaluation of large language models by sparse optimization. In *International Conference on Learning Representations*, 2026.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *International Conference on Artificial Intelligence and Law*, 2021.

Jin Zhou, Christian Belardi, Ruihan Wu, Travis Zhang, Carla Gomes, Wen Sun, and Kilian Q. Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations*, 2025.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, et al. General scales unlock ai evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026. doi: 10.1038/s41586-026-10303-2.

A Positioning Against Adjacent Model-Evaluation Paradigms

B Claim-to-Evidence Validity Card

Table 5 states the map of Section 5 in the reporting form proposed in the Discussion. Each row pairs the facet being varied, the formal object of Section 3.2 or 3.4 that the facet stresses, the study that varies it, and the decision resolution the evidence supports.

The hierarchical sensitivity analysis resolves dependence of the primary contrast on the operational target taxonomy (Appendix H). Broader validity still requires measurement of all claim levels within a single frozen cohort and replication in an independently constructed atlas.

Table 4: CapAtlas complements sparse score prediction with an external-validity protocol. "Independent oracle" means that target outcomes are excluded from map construction and opened only after placement.

Line of work	Sparse target	Independent oracle	Unseen lineage	Local peers	Claim-wise validity
Model Atlas / behavioral maps	Varies	No	Varies	Yes	No
tinyBenchmarks / metabench	Yes	No	Varies	No	No
General Scales	No	Yes	Partial	No	No
DKPS	Yes	No	Partial	Yes	No
CapAtlas	Yes	Yes	Yes	Yes	Yes

Table 5: Validity card. Each facet is paired with the formal object it stresses, the study that varies it, and the resolution the resulting evidence supports.

Facet	Formal object	Validation study	Supported resolution
Population	Supervised pseudometric (Eq. 1)	70 targets / 43 new lineages	Graded local retrieval supported under the primary ontology
Construct	Top-K margin condition (Prop. 1)	Two disjoint ontologies	Reference-only refitting recovers global order; exact local Top-5 remains unresolved
Interface	Fixed fingerprint ϕ_S	Budget, reselection, prompts	Sparse panels work; panel identity and format matter
Mechanism	Conditional-mean capability chart	PC1, matching, nuisance controls	Beyond scalar ability, but matched by calibration-only at the aggregate resolution
Reliability	Rank–precision separation (Prop. 2)	Frozen 50-target selector	AUROC .731; 80.95% precision at 42.0% coverage
Deployment	Fixed-precision acceptance	Five support positives; triage	Defines sample-size and method-specificity targets

C Detailed Cohort Provenance

Table 6: Full-checkpoint cohorts. Codes name archived protocol directories and are used in the appendix audit trail. The frozen cohort has an externally timestamped protocol and documented post-freeze execution amendments; earlier cohorts are development or audited supporting evidence.

Code	Role	Models	Families
P0c	Reference atlas, development half	23	7
P0d	Reference atlas, replication half	23	17
P0f	Development targets, validation	20	16
P0k	Development targets, secondary holdout	20	12
P0n	Family-disjoint hard-baseline confirmation	18	18
P0u	Corrected retrospective audit	20	15
P0w	First frozen lineage-disjoint confirmation	31	20 lineages
P0y	Final frozen replication and extension	70	43 lineages
P0ad	Frozen repeated-target panel audit	26	26 lineages
P0ae	Benchmark-disjoint ontology refitting	70	43 lineages
P0ag	Prospective fixed-precision test	50	45 lineages

The primary cohort of Section 4.1 is P0y; the independent-construct study is P0ae, which reuses the P0y targets and responses under O_2; and the prospective study is P0ag under O_3. The staged design progressively isolates the transferable signal. Independent replication redirected the project from disagreement acquisition to fixed panels; family-disjoint confirmation established supervised decoding over hard-answer matching; and a strict leave-family-out mechanism study identified a compressible soft signal and selected the rank-4 candidate. An internal oracle-representation audit preserved the original source and outputs while classifying the corrected results as retrospective. The first prospective cohort established the SoftMap-versus-soft-JS effect, and the larger final cohort froze the two strongest alternative geometries and an open-set test before evaluation. Target inclusion in both frozen cohorts uses structural completion checks rather than result values.

The P0c–P0y counts refer to unique checkpoints within the principal staged analysis. P0ad and P0ae are listed as separate evaluations but reuse targets from earlier cohorts and must not be added to a unique-checkpoint total. P0ag contains the 26 P0ad checkpoints plus 24 frozen additions; its prospectively sealed evidence consists of the independent-ontology outcomes. The two atlas

halves contain 23 checkpoints from seven declared families and 23 from 17; the validation and secondary-holdout development cohorts contain 20 from 16 and 20 from 12; the family-disjoint confirmation contains 18 from 18; the corrected-audit cohort contains 20 from 15; the first frozen cohort contributes 31 checkpoints from 20 model lineages; and the final frozen cohort contributes 70 from 43 further lineages. Families of the family-disjoint confirmation occur in none of its earlier cohorts, and every final-cohort lineage is absent from the 46-model reference atlas and all earlier target cohorts. Corrected-audit checkpoint identities are new, while its family labels may overlap the atlas.

The replication atlas half lost one planned InternLM checkpoint due to tokenizer incompatibility. The validation cohort lost two planned MPT checkpoints due to remote modeling incompatibility. The family-disjoint confirmation completed all 18 frozen targets after documented execution-only repairs. The corrected-audit cohort screened 29 metadata-selected candidates without benchmark questions; 20 executable checkpoints from 15 families were frozen and completed. The frozen cohort froze 34 checkpoints from 23 declared families. Two Zamba2 checkpoints were held out after an unfused fallback made completion procedurally infeasible, and one Bamba checkpoint exceeded 48 GiB under minimum microbatch execution. The frozen Mamba2 checkpoint completed on an 80 GiB A100 with the same manifests and scoring code. The complete-case rule therefore retained 31 checkpoints, 21 declared families, and 20 model lineages after merging the two GLM-4 variants. Padding, kernel, and legacy-cache repairs changed execution compatibility only. The final protocol froze 72 smoke-tested checkpoints from 44 new lineages; two failed on both compute nodes after the allowed recovery, leaving 70 from 43 lineages. A duplicate scheduler launch was detected during execution; original partial files were retained, repeated item identifiers were removed deterministically, and evaluation resumed from unique completed rows. No prompt, panel, oracle, decoder, endpoint, or score-dependent inclusion decision changed. An earlier controlled LoRA experiment is retained as an early precursor and is not included in the principal unique-checkpoint accounting.

D Comparator Specification and Hyperparameters

Table 7 lists every method in Table 1. All supervised fits use the 46 references only; standardization statistics are computed on references. No hyperparameter was tuned on any frozen-cohort target.

Beyond the comparators of Section 4.2, supporting analyses residualize scale, stage, tokenizer, family, and calibration; compare behavioral, nuisance, and capability distances; and vary probe budget, PCA rank, atlas size, and oracle size along frozen scaling axes, in which generic prefix panels differ from the selected panel. The ridge constant $\lambda = 10$ is inherited from the earlier hard-answer ridge decoder and was fixed before the family-disjoint confirmation. The SoftMap rank was selected among $\{1, 2, 4, 6, 8\}$ under leave-family-out folds on the corrected audit cohort before the frozen protocol was written; performance saturates by rank two in the archived sensitivity analysis (Appendix O).

E Primary Panel

Table 8 lists the 16 frozen probes (SHA-256 prefixes of the item identifiers) in selection order. The full identifiers, the reference-only selection code, and the panel hash are in the released package.

F Model Identifiers

Table 9: Audited checkpoint roster. P0y is the final lineage-shift cohort. P0ad is a repeated-target panel audit: nine exact checkpoints overlap the earlier P0w target roster, although none belongs to the 46-model reference atlas.

Key	Model identifier	Declared lineage	Role/status
biomistral7b	BioMistral/BioMistral-7B	mistral	Reference atlas
pythia160m	EleutherAI/pythia-160m	pythia	Reference atlas
llama31_8b_inst	NousResearch/Meta-Llama-3.1-8B-Instruct	llama	Reference atlas
qwen05b_inst	Qwen/Qwen2.5-0.5B-Instruct	qwen	Reference atlas
qwen15b_inst	Qwen/Qwen2.5-1.5B-Instruct	qwen	Reference atlas
qwen3b_inst	Qwen/Qwen2.5-3B-Instruct	qwen	Reference atlas
qwen7b_inst	Qwen/Qwen2.5-7B-Instruct	qwen	Reference atlas
qwen14b_inst	Qwen/Qwen2.5-14B-Instruct	qwen	Reference atlas

Key	Model identifier	Declared lineage	Role/status
qwen_math7b	Qwen/Qwen2.5-Math-7B-Instruct	qwen	Reference atlas
openbio8b	aaditya/Llama3-OpenBioLLM-8B	llama	Reference atlas
olmo7b_base	allenai/OLMo-2-1124-7B	olmo	Reference atlas
olmo7b_dpo	allenai/OLMo-2-1124-7B-DPO	olmo	Reference atlas
olmo7b_inst	allenai/OLMo-2-1124-7B-Instruct	olmo	Reference atlas
olmo7b_sft	allenai/OLMo-2-1124-7B-SFT	olmo	Reference atlas
meditron7b	epfl-llm/meditron-7b	llama	Reference atlas
mistral7b_inst	mistralai/Mistral-7B-Instruct-v0.2	mistral	Reference atlas
openhermes7b	teknium/OpenHermes-2.5-Mistral-7B	mistral	Reference atlas
llama3_8b_inst	NousResearch/Meta-Llama-3-8B-Instruct	llama	Reference atlas
qwen05b_base	Qwen/Qwen2.5-0.5B	qwen	Reference atlas
qwen15b_base	Qwen/Qwen2.5-1.5B	qwen	Reference atlas
deepseek_r1_8b	deepseek-ai/DeepSeek-R1-Distill-Llama-8B	llama	Reference atlas
phi35mini	microsoft/Phi-3.5-mini-instruct	phi	Reference atlas
gemma2_9b	unsloth/gemma-2-9b-it	gemma	Reference atlas
gemma3_4b_inst	unsloth/gemma-3-4b-it	gemma3	Reference atlas
medgemma4b_inst	google/medgemma-4b-it	gemma3	Reference atlas
smollm2_17b_inst	HuggingFaceTB/SmolLM2-1.7B-Instruct	smollm	Reference atlas
tinyllama11b_chat	TinyLlama/TinyLlama-1.1B-Chat-v1.0	tinyllama	Reference atlas
opt13b_base	facebook/opt-1.3b	opt	Reference atlas
bloom17b_base	bigscience/bloom-1b7	bloom	Reference atlas
gptneo13b_base	EleutherAI/gpt-neo-1.3B	gptneo	Reference atlas
falconrw1b_base	tiiuae/falcon-rw-1b	falcon	Reference atlas
stablelm16b_chat	stabilityai/stablelm-2-1_6b-chat	stablelm	Reference atlas
phi2_base	microsoft/phi-2	phi2	Reference atlas
granite31_2b_inst	ibm-granite/granite-3.1-2b-instruct	granite	Reference atlas
qwen_coder15b	Qwen/Qwen2.5-Coder-1.5B-Instruct	qwen	Reference atlas
qwen_math15b	Qwen/Qwen2.5-Math-1.5B-Instruct	qwen	Reference atlas
deepseek_coder13b	deepseek-ai/deepseek-coder-1.3b-instruct	deepseek	Reference atlas
deepseek_math7b	deepseek-ai/deepseek-math-7b-instruct	deepseek	Reference atlas
r1_qwen15b	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	qwen	Reference atlas
yi15_6b_chat	01-ai/Yi-1.5-6B-Chat	yi	Reference atlas
olmo0425_1b_inst	allenai/OLMo-2-0425-1B-Instruct	olmo0425	Reference atlas
saul7b_inst	Equall/Saul-7B-Instruct-v1	mistral	Reference atlas
med42_8b	m42-health/Llama3-Med42-8B	llama	Reference atlas
llama32_3b_inst	unsloth/Llama-3.2-3B-Instruct	llama	Reference atlas
gemma2_2b_inst	unsloth/gemma-2-2b-it	gemma	Reference atlas
mistral7b_v03	mistralai/Mistral-7B-Instruct-v0.3	mistral	Reference atlas
afm45b_base	arcee-ai/AFM-4.5B-Base	afm	P0y complete
afm45b_inst	arcee-ai/AFM-4.5B	afm	P0y complete
aixcoder7b_base	aiXcoder/aixcoder-7b-base	aixcoder	P0y complete
allam7b_inst	ALLaM-AI/ALLaM-7B-Instruct-preview	allam	P0y complete
amber7b	LLM360/Amber	amber	P0y complete
amberchat7b	LLM360/AmberChat	amber	P0y complete
biomedlm27b	stanford-crfm/BioMedLM	biomedlm	P0y complete
calm2_7b	cyberagent/calm2-7b	calm	P0y complete
calm2_7b_chat	cyberagent/calm2-7b-chat	calm	P0y complete
codeshell7b	WisdomShell/CodeShell-7B	codeshell	P0y complete
codeshell7b_chat	WisdomShell/CodeShell-7B-Chat	codeshell	P0y complete
comma7b_2t	common-pile/comma-v0.1-2t	comma	P0y complete
crystalcoder7b	LLM360/CrystalCoder	crystal	P0y complete
ctllm2b_base	m-a-p/CT-LLM-Base	ctllm	P0y complete
danube3_4b_base	h2oai/h2o-danube3-4b-base	danube	P0y complete
danube3_4b_chat	h2oai/h2o-danube3-4b-chat	danube	P0y complete
danube3_500m_base	h2oai/h2o-danube3-500m-base	danube	P0y complete
eurollm17b	utter-project/EuroLLM-1.7B	eurollm	P0y complete
fox16b_base	tensoropera/Fox-1-1.6B	fox	P0y complete
fox16b_inst	tensoropera/Fox-1-1.6B-Instruct-v0.1	fox	P0y complete
galactica67b	facebook/galactica-6.7b	galactica	P0y complete
hunyuan05b_base	tencent/Hunyuan-0.5B-Pretrain	hunyuan	P0y complete
hunyuan18b_inst	tencent/Hunyuan-1.8B-Instruct	hunyuan	P0y complete
hunyuan7b_inst	tencent/Hunyuan-7B-Instruct	hunyuan	P0y complete
hyperclovax05b_inst	naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-0.5B	hyperclovax	P0y complete
index19b_base	IndexTeam/Index-1.9B	index	P0y complete
index19b_chat	IndexTeam/Index-1.9B-Chat	index	P0y complete
instella3b_base	amd/Instella-3B	instella	P0y complete
instella3b_inst	amd/Instella-3B-Instruct	instella	P0y complete
kanana21b_base	kakaocorp/kanana-nano-2.1b-base	kanana	P0y complete
kanana21b_inst	kakaocorp/kanana-nano-2.1b-instruct	kanana	P0y complete
ling_lite	inclusionAI/Ling-lite	ling	P0y complete
llmjp3_18b	llm-jp/llm-jp-3-1.8b	llmjp	P0y complete
llmjp3_37b	llm-jp/llm-jp-3-3.7b	llmjp	P0y complete
llmjp3_37b_inst	llm-jp/llm-jp-3-3.7b-instruct	llmjp	P0y complete
mapneo7b_base	m-a-p/neo_7b	mapneo	P0y complete
mapneo7b_inst	m-a-p/neo_7b_instruct_v0.1	mapneo	P0y complete
megrez3b_inst	Infinigence/Megrez-3B-Instruct	megrez	P0y complete
mgpt13b_large	ai-forever/mGPT-13B	mgpt	P0y complete

Key	Model identifier	Declared lineage	Role/status
mgpt13b_multi	ai-forever/mGPT	mgpt	P0y complete
mimo7b_base	XiaomiMiMo/MiMo-7B-Base	mimo	P0y complete
mimo7b_sft	XiaomiMiMo/MiMo-7B-SFT	mimo	P0y complete
minerva7b_base	sapienzanlp/Minerva-7B-base-v1.0	minerva	P0y complete
minerva7b_inst	sapienzanlp/Minerva-7B-instruct-v1.0	minerva	P0y complete
moonlight16b_a3b	moonshotai/Moonlight-16B-A3B	moonlight	P0y complete
moonlight16b_a3b_inst	moonshotai/Moonlight-16B-A3B-Instruct	moonlight	P0y procedural failure
motif26b	Motif-Technologies/Motif-2.6B	motif	P0y complete
opencoder15b_base	infly/OpenCoder-1.5B-Base	opencoder	P0y complete
opencoder8b_base	infly/OpenCoder-8B-Base	opencoder	P0y complete
opencoder8b_inst	infly/OpenCoder-8B-Instruct	opencoder	P0y complete
orion14b_chat	OrionStarAI/Orion-14B-Chat	orion	P0y complete
palmyra_base	Writer/palmyra-base	palmyra	P0y complete
palmyra_small	Writer/palmyra-small	palmyra	P0y complete
pleias12b	PleIAs/Pleias-1.2b-Preview	pleias	P0y complete
pleias350m	PleIAs/Pleias-350m-Preview	pleias	P0y complete
pleias3b	PleIAs/Pleias-3b-Preview	pleias	P0y complete
polylm17b	DAMO-NLP-MT/polylm-1.7b	polylm	P0y complete
powerlm3b	ibm/PowerLM-3b	powerlm	P0y complete
powermoe3b	ibm/PowerMoE-3b	powerlm	P0y complete
salamandra2b	BSC-LT/salamandra-2b	salamandra	P0y complete
salamandra7b	BSC-LT/salamandra-7b	salamandra	P0y complete
salamandra7b_inst	BSC-LT/salamandra-7b-instruct	salamandra	P0y complete
sarashina2_7b	sbintuitions/sarashina2-7b	sarashina	P0y complete
sarvam1_2b	sarvamai/sarvam-1	sarvam	P0y complete
seedcoder8b_base	ByteDance-Seed/Seed-Coder-8B-Base	seed_coder	P0y complete
seedcoder8b_inst	ByteDance-Seed/Seed-Coder-8B-Instruct	seed_coder	P0y complete
skywork13b_base	Skywork/Skywork-13B-base	skywork	P0y complete
teuken7b_inst	openGPT-X/Teuken-7B-instruct-research-v0.4	teuken	P0y complete
trillion7b	trillionlabs/Trillion-7B-preview	trillion	P0y complete
viking7b	LumiOpen/Viking-7B	viking	P0y complete
xlstm7b	NX-AI/xLSTM-7b	xlstm	P0y procedural failure
yuan2_2b	IEITYuan/Yuan2-2B-hf	yuan	P0y complete
apertus8b_base	swiss-ai/Apertus-8B-2509	apertus	P0ad; new checkpoint
exaone4_1b2	LGAI-EXAONE/EXAONE-4.0-1.2B	exaone4	P0ad; new checkpoint
nemotron_mini4b	nvidia/Nemotron-Mini-4B-Instruct	nemotron_mini	P0ad; new checkpoint
minicpm4_0b5	openbmb/MiniCPM4-0.5B	minicpm4	P0ad; repeats P0w checkpoint
biogpt_base	microsoft/biogpt	biogpt	P0ad; repeats P0w checkpoint
lfm2_1b2	LiquidAI/LFM2-1.2B	lfm2	P0ad; new checkpoint
mamba1b4	state-spaces/mamba-1.4b-hf	mamba1	P0ad; repeats P0w checkpoint
pharia7b	Aleph-Alpha/Pharia-1-LLM-7B-control-hf	pharia1	P0ad; new checkpoint
codegen2_1b	Salesforce/codegen2-1B	codegen2	P0ad; new checkpoint
mamba2_1b3	AntonV/mamba2-1.3b-hf	mamba2	P0ad; repeats P0w checkpoint
falcon_h1_1b5	tiiuae/Falcon-H1-1.5B-Base	falcon_h1	P0ad; new checkpoint
bamba9b	ibm-ai-platform/Bamba-9B	bamba	P0ad; new checkpoint
zamba2_1b2	Zyphra/Zamba2-1.2B	zamba2	P0ad; new checkpoint
polycoder2b7	NinedayWang/PolyCoder-2.7B	polycoder	P0ad; repeats P0w checkpoint
glm4_9b	THUDM/glm-4-9b-0414	glm4	P0ad; repeats P0w checkpoint
falcon_mamba7b	tiiuae/Falcon-Mamba-7B	falcon_mamba	P0ad; new checkpoint
gptj6b	EleutherAI/gpt-j-6b	gptj	P0ad; repeats P0w checkpoint
gpt_oss20b	openai/gpt-oss-20b	gpt_oss	P0ad; new checkpoint
olmoe_1b7b	allenai/OLMoE-1B-7B-0125	olmoe	P0ad; new checkpoint
rwkv4_1b5	RWKV/rwkv-4-1b5-pile	rwkv4	P0ad; new checkpoint
granite4h_1b	ibm-granite/granite-4.0-h-1b	granite4_hybrid	P0ad; new checkpoint
persimmon8b	adept/persimmon-8b-base	persimmon	P0ad; new checkpoint
starcoder2_3b	bigcode/starcoder2-3b	starcoder2	P0ad; repeats P0w checkpoint
qwen3_0b6_base	Qwen/Qwen3-0.6B-Base	qwen3	P0ad; new checkpoint
incoder1b	facebook/incoder-1B	incoder	P0ad; repeats P0w checkpoint
redpajama3b	togethercomputer/RedPajama-INCITE-Base-3B-v1	redpajama_incite	P0ad; new checkpoint

G Primary-Cohort Secondary Endpoints

Table 10 reports every prespecified secondary metric for the existing-data P0y-R analysis. Relative to the authoritative frozen artifact, method-level NDCG values in this separately hashed recomputation differ by at most 0.0015 (including −0.00021 for SoftMap, −0.00072 for soft-JS, −0.0015 for metric

Table 7: Inputs, model class, hyperparameters, and provenance of every comparator. "Soft" is the 64-dimensional concatenation of four-way probe probabilities; "hard" is the one-hot answer; z is the standardized eight-coordinate reference profile.

Method	Inputs	Estimator and constants	Status
SoftMap	soft	PCA (4 comp., reference-only, unscaled scores) then multi-output ridge $\lambda = 10$	Frozen primary
Soft-JS	soft	Mean itemwise Jensen–Shannon distance; nearest-profile estimate for RMSE	Frozen baseline
Full soft ridge	soft	Ridge $\lambda = 10$ on 64 dims	Frozen
Hard ridge	hard	Ridge $\lambda = 10$ on 64 one-hot dims	Frozen
PLS, rank 4	soft (standardized)	PLS regression, 4 components	Frozen
CCA, rank 4	soft (standardized)	CCA to 4 canonical variates, then ridge $\lambda = 10$ to z	Frozen
BenchPress-style rank 2	soft + z block	Iterative rank-2 SVD completion of the stacked [features, z] matrix, 100 iterations, target z block missing	Frozen
Metric learning	soft pairs	Non-negative diagonal Mahalanobis weights fit by ridge $\lambda = 10$ on squared feature differences vs. squared z distances over all reference pairs; estimate is the mean z of 5 nearest references	Frozen
PhyloLM-style	soft	Nei genetic distance over probes, non-metric MDS to 15 dims (joint per target), MLP (200 ReLU units, Adam 10^{-3}) from embedding to z	Frozen
Metadata only	metadata	Ridge $\lambda = 10$ on log parameters, log vocabulary, stage, model-type and reference-family indicators	P0y-R retrospective
Parameter count	log parameters	Ridge $\lambda = 10$	P0y-R retrospective
Scalar ability	16-item accuracy	Ridge $\lambda = 10$	Frozen
Calibration only	calibration	Ridge $\lambda = 10$ on mean entropy, mean max probability, and four mean answer-option probabilities	P0y-R retrospective
Nuisance only	metadata + calibration	Ridge $\lambda = 10$ on standardized log parameters, log vocabulary, stage, reference-family and model-type indicators, mean entropy, mean max probability, and four mean answer-option probabilities	P0y-R retrospective
DKPS (OLS, 1-NN, $\sqrt{N}$-NN)	soft	Classical MDS (8 dims) of Frobenius distances on references plus one target, then OLS to z, or mean z of 1 or $\lceil\sqrt{46}\rceil = 7$ nearest references	Post-freeze
DKPS cross-task ensemble	soft + accuracy	α SampleScore-OLS $+(1-\alpha)$ DKPS-OLS with $\alpha = 16/240$	Post-freeze

Table 8: The frozen static-16 panel: four items per domain, ordered by round-robin Fisher information.

Item (prefix)	Domain	MMLU subject	Selection rank
7d4ae265	medicine	professional_medicine	1
85ebb87a	law	jurisprudence	2
aa21d8c0	cs	computer_security	3
140c2fcb	math	abstract_algebra	4
29374283	medicine	clinical_knowledge	5
3c01a8cb	law	jurisprudence	6
e3d39d9f	cs	college_computer_science	7
df2f4c2f	math	high_school_mathematics	8
8279ff0a	medicine	medical_genetics	9
9d87c980	law	professional_law	10
680a6bce	cs	college_computer_science	11
9093b7b6	math	high_school_statistics	12
51f2772d	medicine	clinical_knowledge	13
478248b5	law	jurisprudence	14
7a6dada4	cs	computer_security	15
174266b0	math	high_school_statistics	16

learning, and −0.0010 for DKPS-$\sqrt{N}$-NN). We retain the authoritative values for the frozen primary claim and use the recomputation only for retrospective analyses. The provenance ledger identifies both files and quarantines the older analysis-tree output.

H Hierarchical-Lineage Sensitivity

Table 11 reaggregates the unchanged P0y target scores at four outcome-independent resolutions. The first is an unclustered checkpoint analysis; the second is the registered declared-lineage estimand; the last two are reproducible metadata coarsenings. All four satisfy the same positive-interval, $p \le .05$, and 60% sign-majority rule. Holm adjustment across the four sensitivity tests leaves every result significant ($p_{\text{Holm}} \le .0166$).

Table 10: P0y-R lineage-macro secondary endpoints on 70 targets from 43 lineages. Lower is better for regret and RMSE; all analyses are retrospective and cannot alter the frozen P0y decision.

Method	NDCG@5 ↑	Recall@5 ↑	Kendall τ ↑	Regret ↓	RMSE ↓
SoftMap	.8607	.4271	.6597	.7615	.7773
Full soft ridge	.8586	.4124	.6652	.7246	.7738
Hard ridge	.8345	.4023	.6070	.9615	.8359
Scalar ability	.7634	.1736	.4935	1.2329	.9456
PLS-4	.8728	.4907	.6646	.7754	.7712
Metric learning	.8678	.4814	.6526	.7988	.7798
DKPS-OLS	.8672	.4767	.6606	.8264	.8044
DKPS-1NN	.8258	.4031	.5817	1.1337	.9353
DKPS-$\sqrt{N}$NN	.8573	.4612	.6313	.8744	.8073
DKPS ensemble	.8663	.4775	.6609	.8264	.7981
Parameter count	.7005	.1682	.2946	1.5743	1.0636
Metadata only	.7725	.2411	.4749	1.2622	.9341
Calibration only	.8611	.4442	.6417	.8102	.7980
Nuisance only	.8646	.4775	.6448	.8698	.8120
Soft-JS	.8174	.3853	.4155	1.1687	.9477

Table 11: Hierarchical sensitivity of the frozen SoftMap–soft-JS NDCG@5 contrast. Each row macro-averages equally over the indicated grouping. Publisher namespace and implementation class are reproducible coarsenings, not claims of canonical ancestry. Intervals are 100,000-draw group bootstraps; p values are two-sided sign flips.

Grouping	Groups	Δ NDCG@5	95% interval	Positive	p
Checkpoint	70	.0366	[.0192, .0544]	47/70	1.09×10^{-4}
Declared release lineage	43	.0428	[.0228, .0641]	31/43	1.14×10^{-4}
Publisher namespace	41	.0429	[.0235, .0642]	30/41	9.00×10^{-5}
Implementation architecture	20	.0356	[.0098, .0606]	14/20	.0166

We separately preserved equal declared-lineage weighting and changed only the dependence unit. Publisher-cluster inference gives 0.0428 ([0.0216, 0.0652], sign-flip $p = 0.000180$). Because two declared lineages cross frozen `model_type` labels, the architecture analysis uses the 18 connected components of the declared-lineage–implementation bipartite graph; it gives 0.0428 ([0.0169, 0.0554], exact $p = 0.00835$). Finally, all 43 leave-one-lineage-out analyses and all 260 single merges allowed by a shared publisher or implementation label were enumerated. The worst deletion remains 0.0372 ([0.0193, 0.0554], $p = 0.000199$), and the worst merge remains 0.0384 ([0.0200, 0.0577], $p = 0.000233$). Thus neither one influential lineage nor one disputed split explains the primary effect. Publisher and implementation labels are sensitivity groupings, not asserted canonical ancestry.

I Independent Capability Construct: Detail

Table 12 reports all method-level secondary endpoints for the reference-only refitting under O_2 (P0ae). For SoftMap minus soft-JS, the paired lineage effects are: Recall@5 +0.0008 ([−0.0930, 0.0992]), Kendall's τ +0.1691 ([0.1335, 0.2061]), top-1 regret +0.0114 ([−0.1317, 0.1490]; lower is better), and capability RMSE −0.0144 ([−0.0697, 0.0379]). For SoftMap minus calibration-only they are −0.0155, −0.0162, −0.0135, and −0.0383, respectively; all intervals cross zero. Only the SoftMap-versus-soft-JS Kendall interval excludes zero; no secondary endpoint enters the prespecified familywise decision. The machine-readable result and frozen manifest hashes are recorded in the provenance ledger.

We additionally audited whether the thinner second-ontology oracle could itself produce the apparent resolution difference. Even–odd splits within each task (24 items per half) give bidirectional lineage-macro NDCG@5 0.9385, Recall@5 0.6333, and Kendall's $\tau = 0.8155$. The oracle distance-vector agreement is 0.8155 ([0.7948, 0.8345]), whereas Top-5 neighbor Jaccard is 0.4968 ([0.4343, 0.5615]) and top-1 agreement is 0.2481 ([0.1434, 0.3682]). In 20,000 nested bootstrap draws that resample items within tasks and then lineages, the SoftMap-minus-soft-JS Kendall effect is 0.1731 ([0.1324, 0.2145]), while the NDCG@5 effect is 0.0112 ([−0.0398, 0.0639]). A direct global-minus-within-Top-10 pair-accuracy contrast is 0.0419 ([−0.0132, 0.0936]) and remains descriptive. These retrospective checks support robust coarse ordering and limited local-oracle resolution; they do not convert the failed frozen co-primary decision into a success or establish equivalence at the local endpoint.

J Third-Ontology Oracle Reliability

We applied the same deterministic even–odd audit to the prospective O_3 outcomes, using 32 items per task per half and recomputing each half's standardization from the 46 references. Across 45 target lineages, bidirectional split-half NDCG@5 is 0.9672 ([0.9558, 0.9774]), Recall@5 is 0.7489 ([0.6822, 0.8111]), Top-5 Jaccard is 0.6440 ([0.5664, 0.7198]), and distance-vector Kendall's τ is 0.8676 ([0.8444, 0.8887]). Top-1 identity is the least stable endpoint at 0.4333 ([0.2889, 0.5778]), consistent with the resolution-specific account.

For the certificate's direct target, we held the frozen predicted Top-5 fixed and recomputed whether Recall@5 reached 0.6 under each oracle half. The two resolved/unresolved labels agree for 0.9111 of lineages ([0.8222, 0.9778]); each half agrees with the full-oracle label for 0.9556 ([0.8889, 1.0000]). These retrospective reliability estimates leave the prospective selector, threshold, and FAIL decision unchanged. They show that the registered precision gap is measured against a reproducible binary target, while exact neighbor identity remains the finer and less stable object. Holding the frozen 21-target accepted set fixed, half A resolves 17 targets (80.95%) and half B resolves 18 (85.71%); both remain below 90%, and only one accepted target changes label between halves.

Table 12: P0ae benchmark-disjoint ontology: lineage-macro endpoints on the unchanged 70 targets.

Method	NDCG@5	Recall@5	Kendall τ	Regret	RMSE
SoftMap	0.8310	0.4481	0.7640	0.4449	0.3223
Full soft ridge	0.8355	0.4512	0.7673	0.4248	0.3146
Scalar ability	0.6089	0.0907	0.4471	0.9636	0.5665
Calibration only	0.8372	0.4636	0.7802	0.4584	0.3606
Soft-JS	0.8230	0.4473	0.5949	0.4335	0.3367

K Prospective Reliability Cases

The qualitative cases in Figure 4 are selected by outcome cell rather than by model identity: accepted–resolved, rejected–unresolved, and accepted–unresolved. The complete frozen cohort is listed below in its original roster order.

Table 13: Complete P0ag target ledger in frozen roster order. The three Figure 4 cases represent accepted–resolved, rejected–unresolved, and accepted–unresolved outcome cells; all 50 targets are shown here.

Target	Lineage	Score	Accept	Recall@5	NDCG@5	Outcome
apertus8b_base	apertus	0.305	no	0.2	0.693	unresolved
exaone4_1b2	exaone4	0.073	no	0.0	0.581	unresolved
nemotron_mini4b	nemotron_mini	0.625	yes	0.4	0.903	unresolved
minicpm4_0b5	minicpm4	0.457	no	0.2	0.821	unresolved
biogpt_base	biogpt	0.120	no	1.0	0.961	resolved
lfm2_1b2	lfm2	0.342	no	0.4	0.833	unresolved
mamba1b4	mamba1	0.593	yes	0.6	0.934	resolved
pharia7b	pharia1	0.071	no	0.0	0.641	unresolved
codegen2_1b	codegen2	0.782	yes	0.8	0.992	resolved
mamba2_1b3	mamba2	0.782	yes	1.0	0.999	resolved
falcon_h1_1b5	falcon_h1	0.418	no	0.0	0.692	unresolved
bamba9b	bamba	0.277	no	0.0	0.795	unresolved
zamba2_1b2	zamba2	0.763	yes	0.2	0.791	unresolved
polycoder2b7	polycoder	0.193	no	0.8	0.993	resolved
glm4_9b	glm4	0.225	no	0.6	0.938	resolved
falcon_mamba7b	falcon_mamba	0.509	no	1.0	0.968	resolved
gptj6b	gptj	0.714	yes	0.8	0.985	resolved
gpt_oss20b	gpt_oss	0.419	no	0.0	0.764	unresolved
olmoe_1b7b	olmoe	0.633	yes	0.4	0.825	unresolved
rwkv4_1b5	rwkv4	0.842	yes	0.8	0.992	resolved
granite4h_1b	granite4_hybrid	0.111	no	0.2	0.841	unresolved
persimmon8b	persimmon	0.800	yes	0.8	0.977	resolved
starcoder2_3b	starcoder2	0.595	yes	0.8	0.984	resolved
qwen3_0b6_base	qwen3	0.625	yes	0.4	0.831	unresolved
incoder1b	incoder	0.335	no	1.0	0.999	resolved
redpajama3b	redpajama_incite	0.465	no	0.6	0.916	resolved
decilm7b	decilm	0.061	no	0.4	0.937	unresolved
exaone35_2b4	exaone35	0.220	no	0.4	0.886	unresolved
exaone35_7b8	exaone35	0.333	no	0.6	0.928	resolved

Continued on next page

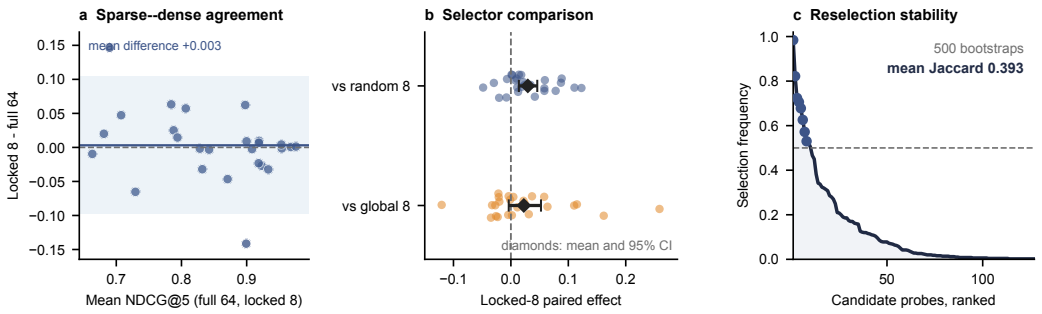


Figure 5: A sparse interface preserves dense-panel performance while revealing panel specificity. (a) Lineage-level agreement between locked eight-probe and full 64-item interfaces; vertical segments show deviations from identity. (b) Locked-8 exceeds random-8 and performs comparably to global-8; diamonds show paired means and descriptive 95% intervals. (c) Five hundred reference-lineage bootstraps quantify probe-selection variability.

Table 13 – continued

Target	Lineage	Score	Accept	Recall@5	NDCG@5	Outcome
bitnet2b_bf16	bitnet	0.507	no	0.2	0.791	unresolved
gpt2_small	gpt2	0.206	no	1.0	0.926	resolved
gpt2_medium	gpt2	0.239	no	0.8	0.923	resolved
gpt2_xl	gpt2	0.759	yes	0.8	0.983	resolved
pythia2b8	pythia	0.761	yes	1.0	0.999	resolved
redpajama3b_inst	redpajama_incite	0.807	yes	1.0	0.998	resolved
granite31moe3b	granite31_moe	0.416	no	0.6	0.945	resolved
hunyuan7b_pre	hunyuan_v1	0.216	no	0.2	0.833	unresolved
internlm25_1b8	internlm25	0.549	no	0.8	0.980	resolved
numinamath7b	deepseek_math	0.528	no	0.2	0.860	unresolved
smollm2_360m	smollm2	0.808	yes	1.0	0.999	resolved
smollm2_360m_inst	smollm2	0.883	yes	0.6	0.972	resolved
mobillama05b	mobillama	0.858	yes	0.8	0.988	resolved
xgen7b_base	xgen	0.774	yes	0.8	0.945	resolved
falcon3_1b_inst	falcon3	0.210	no	0.6	0.969	resolved
openllama7b_v2	openllama	0.794	yes	0.6	0.974	resolved
minicpm3_4b	minicpm3	0.300	no	0.4	0.818	unresolved
olmo2_1b_sft	olmo2	0.584	yes	0.6	0.965	resolved
openelm1b1_inst	openelm	0.456	no	0.0	0.622	unresolved
phi1b5	phi15	0.463	no	0.2	0.894	unresolved
plamo2_1b	plamo2	0.704	yes	1.0	0.990	resolved

L Panel and Interface Stability

The repeated-target panel audit (P0ad) uses 26 checkpoints from 26 declared lineages, all absent from the reference atlas, although nine exact checkpoint IDs overlap P0w. Every target completed 64 probe items and the 1,564-item oracle. Locked8 was fixed in an earlier analysis, global8 used the eight most disagreeing reference items, random8 averaged 200 shared random panels, and full64 used the complete pool; all used unweighted soft-JS. Their NDCG@5 values are 0.8510, 0.8285, 0.8215, and 0.8478, respectively. Locked8 minus random8 is +0.02947 ($[0.01387, 0.04593]$; 21/26 positive lineages), whereas locked8 minus global8 is +0.02250 ($[-0.00373, 0.05242]$; 14/26 positive). The audit therefore establishes sparse-panel stability and a dense-panel ratio of 1.004; the registered joint criterion remains unmet because the global8 contrast is uncertain. Figure 5 summarizes the interface comparison and the reselection variability discussed in Section 6.

M Benchmark Triage Detail

Given a predicted capability vector and a task-weight vector, the auditor can shortlist a reference model before paying for full evaluation. After each acquired benchmark coordinate, that coordinate is replaced by its observation. Uncertainty radii are calibrated from leave-reference-lineage-out residuals; the next coordinate greedily reduces the resulting utility-regret bound. The policy stops only when this target-label-blind upper bound is at most 10% of the corresponding reference-atlas

utility range. We report empirical violations among certified portfolios and the fraction of eight equal-size benchmark coordinates not acquired. The stopping rule never inspects an unacquired target oracle value. Triage evaluates eight single-task utilities, 28 task pairs, and 512 task-weight vectors drawn once from Dirichlet($\mathbf{1}$), with and without a parameter-count $\leq$ 4B constraint; logged latency and memory were unavailable for cached evaluations, so nominal memory or latency filters are not interpreted.

Table 14: Outcome-blind benchmark triage over 512 Dirichlet portfolios with parameter count $\leq$ 4B. Radii are calibrated by leave-reference-lineage-out residuals. Violation means certified stopping with realized normalized regret above 10%.

Method	Coordinates saved ↑	Violations ↓	Final regret ↓
DKPS, OLS	46.56%	0/512	.00090
SoftMap	46.51%	0/512	.00056
Soft-JS	45.85%	0/512	.00075
Metadata only	43.33%	0/512	.00023
Scalar ability	41.41%	0/512	.00088
Atlas mean	39.87%	0/512	.00085

Final regret values near 10^{-3} show that the reference-calibrated radii are conservative; the rule stops late and uniformly across decoders, which is why savings are nearly method-nonspecific.

N Additional Development Evidence

Acquisition and stopping audits. A selector–decoder factorial audit on the two development target cohorts held decoding fixed while comparing Fisher-information adaptive testing (CAT), static IRT, and balanced random panels. Under the common decoder of that stage, CAT16 minus random16 was +0.0237 (95% CI $[-0.0026, 0.0497]$) on the validation cohort and +0.0262 ($[-0.0050, 0.0561]$) on the secondary holdout. Static16 minus CAT16 was +0.0101 ($[-0.0306, 0.0445]$) and +0.0392 ($[0.0109, 0.0700]$), respectively. CAT16 minus random64 was −0.0146 ($[-0.0337, 0.0056]$) and −0.0325 ($[-0.0635, -0.0035]$). Full240 exceeded CAT16 in both cohorts. A conformal stopping analysis reached 100% coverage at 16 queries, with mean interval width 0.600 and mean maximum width 0.994; every target therefore reached the 64-query cap. These results selected the fixed static panel used from the family-disjoint confirmation onward. Independent replication supplied the deciding evidence: disagreement-based acquisition changed from +0.0294 NDCG over random on the development atlas half to −0.0273 on the independently specified half, and the retrieval-aware variant met none of its frozen gates.

Family-disjoint confirmation and second probe bank. On the identical frozen 16-item panel, direct hard-response matching reached 0.7454 NDCG@5 on 18 unseen families, and ridge reached 0.8540, a family-macro improvement of +0.1086 (97.5% CI $[0.0655, 0.1548]$; exact paired permutation $p = 6.10 \times 10^{-5}$). A structured kernel-likelihood decoder used at that stage reached 0.8273; it was later superseded by SoftMap and is not part of the present method set. With a 16-item panel selected from six non-MMLU datasets using reference responses only, the same contrasts reproduced (+0.1609, $[0.0892, 0.2338]$ for ridge), ruling out an MMLU-specific explanation for the hard-baseline result. Soft-JS was statistically tied with hard ridge on that cohort, which is why the frozen protocol targeted the SoftMap-versus-soft-JS question.

Absolute capability estimation and weight priors. In 24 controlled LoRA training trajectories (96 states), cross-family IRT and expected-information-gain selection established fixed and random panels as the stronger portable baseline for absolute capability estimation. Weight-space summaries produced marked negative transfer in some families. This evidence focused the final capability-atlas method on sparse behavioral fingerprints.

Purpose labels and clustering. Early retrieval experiments evaluated coarse intended-use label transfer and unsupervised clustering. Under leave-family-out at $B = 32$, purpose Macro-F1 was 0.205 and 0.1238 on the two atlas halves; clustering adjusted Rand indices were similarly low. These results narrowed the estimand from metadata completion to capability-relative peer placement.

External-oracle correction. The first external oracle used binary accuracies for all eight tasks and had median split-half distance reliability 0.5997 ([0.4954, 0.6842]) on the validation cohort. The final hybrid oracle replaces multiple-choice correctness with gold-minus-best-wrong likelihood margins while retaining exact-match/execution scores for generative and code tasks. Reliability rises to 0.7407 ([0.6630, 0.8050]) on that cohort and to a median of 0.84 on the frozen cohort. This correction was specified before the secondary-holdout capability responses were collected.

O Post-Freeze Replication and Atlas-Augmentation Analyses

These analyses were run after the frozen analyses on cached data only; no frozen artifact was modified and no new model was evaluated. They are retrospective and are not counted as confirmation.

Cross-cohort replication of the method comparison. Table 15 repeats the full comparison on the four earlier target cohorts with the same 46-model atlas, panel, oracle, and method constants, holding out the target's family from the atlas whenever that family is present in it. The frozen SoftMap-minus-soft-JS contrast is positive in all four cohorts. The square-root-N DKPS advantage observed on the frozen cohort does not appear in any earlier cohort: its paired difference from SoftMap is negative in three and small in the fourth. Metric learning is likewise indistinguishable from SoftMap. The leading methods again sit at or above the split-half ceiling in every cohort.

Table 15: Retrospective replication on earlier target cohorts (family-macro NDCG@5; leave-family-out where the family is in the atlas). Paired contrasts use the exact two-sided family sign-flip test.

Targets / families	Validation (P0f) 18 / 14	Sec. holdout (P0k) 18 / 11	Family-disjoint (P0n) 18 / 18	Corrected audit (P0u) 20 / 15
Split-half ceiling	.879	.874	.873	.875
SoftMap	.853	.881	.871	.879
DKPS, $\sqrt{N}$-NN	.845	.879	.847	.887
Metric learning	.837	.884	.877	.894
Full soft ridge	.851	.880	.871	.881
PLS, rank 4	.855	.860	.884	.890
Soft-JS	.789	.819	.843	.814
Scalar ability	.780	.805	.780	.790
Atlas-mean floor	.740	.752	.749	.729
Random floor	.615	.612	.616	.618
SoftMap − soft-JS	+.064 (p = .031)	+.062 (p = .033)	+.028 (p = .058)	+.066 (p = .029)
DKPS $\sqrt{N}$ − SoftMap	−.009 (p = .75)	−.002 (p = .89)	−.024 (p = .13)	+.008 (p = .65)
Metric − SoftMap	−.016 (p = .59)	+.003 (p = .84)	+.005 (p = .76)	+.014 (p = .62)

Atlas augmentation. Adding to the atlas every completed earlier-cohort checkpoint whose lineage is absent from the frozen cohort (60 checkpoints; 106 references in total) and re-evaluating the 31 frozen targets changes the picture in two ways. First, the SoftMap-minus-soft-JS contrast widens to +0.053 (exact p = 0.002; 15/20 lineages), and SoftMap (0.889) now exceeds square-root-N DKPS (0.879) and metric learning (0.861); absolute NDCG@5 values fall because the task has more candidates and are not comparable with the final-cohort scores in Table 1. Second, the code-lineage failures respond to atlas composition: the polycoder contrast moves from −0.012 to +0.109, starcoder2 from −0.042 to 0.000, and incoder from −0.075 to −0.045, although the augmented atlas adds only one further code-specialized checkpoint. The supervised map benefits from a larger and more diverse atlas more than the transductive neighborhood method does; a prospective test of this observation is part of the released protocol.

Fingerprint-rank sensitivity. The retained fingerprint rank and the capability-space dimension in Proposition 1 are distinct quantities; the following diagnostic does not identify the proposition's ρ_j. Residual capability PCA variance is 0.53, 0.34, 0.21, 0.12, 0.05, 0.03, and 0.01 for ranks one to seven, while model-lineage NDCG@5 for SoftMap at ranks one to eight is 0.906, 0.918, 0.916, 0.917, 0.915, 0.915, 0.913, and 0.912. The retrieval curve saturates by rank two even as residual capability variance continues to fall; this is descriptive evidence of low effective rank, not a test of the sufficient recovery condition.

P Prompt-Format Sensitivity

Every result in the paper uses one canonical plain-text prompt. To measure how much the recovered geometry depends on that interface, we re-scored the frozen 16-item panel for every instruction-tuned checkpoint in the atlas and the frozen cohort whose tokenizer ships a chat template (37 checkpoints; three instruction-tuned checkpoints have no template), wrapping the same content in the model's native template with the instruction to reply with only the option letter. Option letters are scored by summed log-probability exactly as in the frozen probe evaluator. Re-scoring the canonical prompt reproduces the frozen probe rows (median itemwise Jensen–Shannon distance 0.007).

The chat template moves the response distributions substantially: the median itemwise Jensen–Shannon distance between canonical and templated distributions is 0.135 (interquartile range 0.103 to 0.294; maximum 0.431 for a code model), panel accuracy falls by 4.7 points on average, and the mean answer-letter bias shifts by about two percentage points per option. Retrieval is affected in the expected direction. For the three frozen-cohort targets with templates, keeping the atlas canonical and feeding the templated target vectors lowers target-mean NDCG@5 from 0.960 to 0.882 for SoftMap and from 0.830 to 0.729 for soft-JS; switching the instruction-tuned references to their templates as well (a mixed atlas) lowers them further to 0.836 and 0.640. The recovered geometry is therefore interface-relative: probe distributions collected under one prompt format should be compared only with an atlas collected under the same format, and the supervised map degrades more gracefully than direct distribution matching under an interface mismatch. Three targets do not support a statistical claim; the experiment quantifies the sensitivity flagged in the Ethics Statement rather than resolving it.

Q Thick-Oracle Extension

The frozen oracle has 48 items per task, which places the split-half ceiling at 0.893 and leaves the leading methods unresolvable. We subsequently extended every oracle task with 152 additional items drawn from the same source, config, and split by a new stable hash that excludes every frozen item (HumanEval: 116, exhausting its test split), giving 200 items per task (164 for HumanEval). The evaluator, prompts, scoring rule, lengths, model loading flags, atlas, panel, and method constants are unchanged; only the oracle profile is recomputed as the task mean over the enlarged item sets. The extension protocol and its three prespecified contrasts were written before any extension item was evaluated. Two baichuan2 targets did not complete the extension run (their remote modeling code raises a dtype error in the generation path under the current library versions, on two clean attempts) and are excluded as an execution failure, so the analysis covers 29 targets from 19 model lineages.

Table 16: First frozen cohort under the thick oracle (model-lineage macro NDCG@5, 19 lineages). Its original thin-oracle values are repeated for comparison. Paired contrasts use the exact two-sided lineage sign-flip test; Holm-adjusted p in parentheses.

	Thin oracle (48/task)	Thick oracle (200/task)
Split-half ceiling	.893	.960
Metric learning	.928	.942
DKPS, $\sqrt{N}$-NN	.932	.931
SoftMap	.917	.929
Full soft ridge	.911	.926
PLS, rank 4	.891	.912
DKPS, 1-NN	.892	.906
DKPS, OLS	.890	.896
Hard ridge	.874	.891
Soft-JS	.882	.889
Scalar ability	.776	.763
Atlas-mean floor	.698	.640
Random floor	.594	.582
SoftMap − soft-JS	+.034 (p = .028)	+.040 (p = .028; Holm .085), 13/19
DKPS $\sqrt{N}$ − SoftMap	−.016 (p = .22)	+.002 (p = .89), 11/19
Metric − SoftMap	—	+.013 (p = .23), 11/19

Three things change with the thicker oracle (Table 16). The split-half ceiling rises to 0.960, so the leading methods now sit 0.02 to 0.03 below it and their differences are no longer masked by oracle noise. The frozen SoftMap-minus-soft-JS contrast persists (+0.040, exact p = 0.028), although it

does not survive Holm correction over the three contrasts. The thin-oracle advantage of square-root-N DKPS disappears (+0.002, $p = 0.89$), and metric learning's lead (+0.013) is not significant. The two oracles rank the references almost identically for each target (median Spearman correlation between thin and thick oracle distances 0.953), so the thin-oracle results are not an artifact of item choice; they were simply too noisy to separate the leading estimators. The thick oracle is the oracle of the prospective cohort in Appendix R.

R Prospective Cohort: Three-Way Comparison and Enriched Open-Set Test

Protocol. We subsequently froze a second, larger lineage-disjoint cohort to answer three questions with adequate power: whether the frozen SoftMap-versus-soft-JS effect replicates on new lineages under the thick oracle (C1), whether either the transductive DKPS geometry or metric learning is better than the supervised map (C2, C3), and whether any retrieval geometry can flag out-of-support targets (open set). Candidates were 114 checkpoints from 62 declared lineages absent from the atlas and from every earlier cohort, with an a-priori *out-of-support candidate* flag (non-English-centric pretraining, fewer than 150M parameters, or a narrow specialty corpus) written before evaluation. After neutral download and smoke preflight, 72 checkpoints from 44 lineages were frozen with the protocol, panel hash, and both oracle manifests under an RFC 3161 timestamp (2026-09-06 00:46 UTC); the candidate list, exclusions (20 gated or unavailable repositories, 21 smoke failures, one previously evaluated model id), and every execution-only amendment are in the released package. The oracle is the 200-item thick oracle of Appendix Q; methods and constants are unchanged. The three contrasts are exact lineage sign-flip tests (Monte Carlo, 10^6 signs) with Holm correction; a method is declared better only with Holm $p \leq 0.05$ and a 60% sign majority. Two frozen checkpoints (moonlight16b_a3b_inst, xlstm7b) failed on both machines after the allowed retry and are excluded as execution failures, leaving 70 targets from 43 lineages; 36 targets were evaluated on a second machine (A100, torch 2.5.1) with the same code and library versions otherwise.

Table 17: Prospective cohort under the thick oracle (lineage-macro NDCG@5 over 43 lineages; 70 targets; 46-model atlas). Contrasts are prespecified.

Method	NDCG@5
Split-half oracle ceiling	.956
PLS, rank 4	.873
Metric learning	.869
DKPS, OLS	.867
DKPS, cross-task ensemble	.866
SoftMap	.861
Full soft ridge	.859
DKPS, $\sqrt{N}$-NN	.858
Hard ridge	.834
DKPS, 1-NN	.826
Soft-JS	.818
Scalar ability	.763
Atlas-mean floor	.687
Random floor	.605
C1: SoftMap − soft-JS	+.0428; $p = .000125$ (Holm .000375); 31/43 positive; declared better
C2: DKPS $\sqrt{N}$ − SoftMap	−.0026; $p = .740$ (Holm .740); 21/43; not distinguishable
C3: Metric − SoftMap	+.0084; $p = .337$ (Holm .673); 18/43; not distinguishable

Three-way result. The frozen effect replicates prospectively: SoftMap exceeds same-input soft-JS by 0.0428 lineage-macro NDCG@5 on 43 unseen lineages, with 31 positive lineages and Holm-adjusted $p = 3.75 \times 10^{-4}$ (descriptive lineage bootstrap $[0.0229, 0.0640]$). Neither the transductive nearest-neighbor geometry nor metric learning is distinguishable from the supervised map (C2 −0.0026, C3 +0.0084), and the descriptive spread among all supervised and transductive estimators (0.857 to 0.873) is smaller than their common gap to direct matching. Absolute scores are lower than on the earlier frozen cohort of Appendix Q (0.86 versus 0.93) although the ceiling is similar (0.956): the new lineages lie farther from the atlas, and every method loses accuracy together, which is the pattern expected when atlas coverage rather than estimator choice limits retrieval (Appendix O).

Open set. Ground truth labels a target out of support when its nearest atlas reference in thick-oracle space is farther than the 95th percentile of leave-one-out nearest-reference distances among the

references (threshold 2.69 standardized units). Only five of 70 targets meet it, all code-specialized or Chinese mixture-of-experts models (aixcoder, seed-coder, opencoder-instruct, ling, megrez); none of the 22 a-priori flagged non-English or tiny checkpoints does (Fisher exact $p = 0.17$). The enrichment hypothesis was therefore wrong in an informative way: multilingual pretraining moved targets within the atlas hull, whereas code specialization moved them outside it, consistent with the code-lineage failures analyzed in Appendix O. With five positives the prespecified support criterion (at least eight positives and a bootstrap AUROC lower bound above 0.5) cannot be met for any detector; descriptively, the distance from the SoftMap estimate to its nearest reference has AUROC 0.76 (95% bootstrap $[0.48, 0.98]$), metric-learning distance 0.63, DKPS embedding distance 0.505, and minimum soft-JS 0.42. The paper's statement that open-set certification is unsupported stands, now with a prospective negative result and a concrete lead (code specialization) for the next enrichment.

S Theory Details and Proofs

S.1 Population target and nuisance leakage

Equation 1 is a pullback pseudometric: non-negativity, symmetry, and the triangle inequality follow from the Euclidean metric, while distinct fingerprints may have zero distance when their conditional means agree. For centered fingerprints $\mathbf{x}$ and profiles $\tilde{\mathbf{z}}$, the corresponding population objective with penalty $\bar{\lambda}$ has solution

$$M^*_{\bar{\lambda}} = \Sigma_{zx}(\Sigma_{xx} + \bar{\lambda} I)^{-1}. \tag{5}$$

Under $\mathbf{x} = A\tilde{\mathbf{z}} + H\mathbf{n} + \boldsymbol{\epsilon}$,

$$\Sigma_{zx} = \Sigma_z A^\top + \Sigma_{zn} H^\top + \Sigma_{z\epsilon}. \tag{6}$$

Thus even with mean-independent noise, nuisance enters the optimal linear chart whenever $\Sigma_{zn} \neq 0$. The method targets conditional-mean capability; nuisance-only and calibration-only predictors quantify the shared aggregate structure, while residualization measures its dependence on representation. The observed estimator band is consistent with several plug-ins approximating the same low-dimensional target and motivates richer estimators of m_O.

S.2 Proof of Proposition 1

For every reference j, the reverse and ordinary triangle inequalities give

$$d_z(t, j) - \delta_t \leq \|\hat{\mathbf{z}}_t - \tilde{\mathbf{z}}_j\| \leq d_z(t, j) + \delta_t. \tag{7}$$

Every $i \in T_K(t)$ therefore has predicted distance at most $d_{(K)} + \delta_t$. Equation 3 makes every $j \notin T_K(t)$ exceed that value, so no outside candidate can displace a true neighbor. This proves exact set recovery. More generally, let

$$V_t = \{j \notin T_K(t) : d_z(t, j) - d_{(K)} \leq 2\delta_t\}.$$

Candidates in V_t define the complete set that can cross the boundary, giving Recall@ $K \geq 1 - \min\{K, |V_t|\}/K$. This worst-case guarantee connects estimation error to the attainable neighborhood resolution of an observed atlas.

The prediction error can be related to the fingerprint model used for interpretation. For centered variables, write $\phi_t = A\tilde{\mathbf{z}}_t + H\mathbf{n}_t + \boldsymbol{\epsilon}_t$ and $M = \hat{W}P_4$. Then

$$\delta_t \leq \|(MA - I)\tilde{\mathbf{z}}_t\| + \|M(H\mathbf{n}_t + \boldsymbol{\epsilon}_t) + \hat{\mathbf{b}}\|. \tag{8}$$

The first term is capability distortion, including directions discarded by the rank-four chart; the second collects nuisance, response noise, and intercept error. Equation 8 identifies the error components that must be controlled to preserve a narrow Top-K boundary.

S.3 Proof and instantiation of Proposition 2

Let $u = \Pr(R \geq \tau \mid Y = 0)$ and $v = \Pr(R \geq \tau \mid Y = 1)$ be the false- and true-positive rates. Then $\kappa = \theta v + (1 - \theta)u$ and $\gamma = \theta v/\kappa$, hence $u = \kappa(1 - \gamma)/(1 - \theta)$ and $v = \gamma\kappa/\theta$. Any nondecreasing ROC curve through (u, v) has area at most

$$uv + (1 - u) = 1 - u(1 - v),$$

because it is at most v to the left of u and at most one to the right. Substitution proves Equation 4. AUROC equals $\Pr(R^+ > R^-)$ plus half the tie probability, so it is unchanged by a strictly increasing transformation. Its ROC curve is also unchanged by pure class-prior shift when the class-conditional score laws remain fixed, whereas

$$\mathrm{Prec}_\theta(\tau) = \frac{\theta v}{\theta v + (1-\theta)u} \tag{9}$$

varies with θ whenever $u, v > 0$. This establishes the non-implication.

For the certification cohort, $a = 0.731$, $\kappa = 21/50 = 0.42$, and $\theta = 29/50 = 0.58$. Equation 4 reduces to

$$\frac{21}{29}\gamma^2 - \frac{50}{29}\gamma + 0.731 \le 0,$$

whose relevant root is $\gamma = 0.5519$. Precision one is also compatible with these summaries: setting $u = 0$ gives $v = \kappa/\theta = 0.724 < a$, after which a monotone ROC completion can attain area a. Therefore AUROC and coverage restrict precision only to $[0.552, 1]$ here; the registered 0.90 level is neither implied nor excluded.