TOKENPRINT: A CALIBRATED TOKEN-SPACE FINGERPRINT FOR LANGUAGE-MODEL PROVENANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Establishing the provenance of a language model—including its base checkpoint and possible overlap in training distributions—is a governance challenge that metadata alone cannot resolve. We introduce a training-free fingerprint based on the top-k vocabulary projections of late hidden states elicited by 250 fixed knowledge probes, compared using Jaccard overlap over decoded token strings. We evaluate the method on 32 open-weight models from nine families (0.6B–32B) with documented relationships. (1) A *similarity ladder* broadly follows model relatedness: independently trained models on identical data score 0.48 raw (0.35 vocabulary-corrected), followed by shared-base fine-tunes (0.39/0.33), same-developer relatives (0.38/0.28), and models with no documented relationship (0.22/0.17). This identical-data signal persists across three organizations, two tokenizer families, and two architecture classes, and emerges within the first 1% of training before measurable task competence, suggesting a contribution from shared training data beyond capability convergence. (2) As a nearest-neighbor *lineage-retrieval* method, the fingerprint ranks the exact documented base among the top two candidates for all five R1 distillations (mean rank 1.8, MRR 0.60), including a math-specialized base not identifiable from coarse metadata. (3) A *depth ablation* shows that lineage group discrimination strengthens toward the output distribution, with AUC increasing from 0.72 at quarter depth to 0.90 at the output; using only the top 5 output tokens retains AUC 0.87. (4) The fingerprint remains stable under quantization, with Jaccard similarity of 0.92 under int8 and 0.82–0.85 under int4, compared with a maximum cross-model similarity of 0.81 in the calibration pool. We release the probes, code, and fingerprints.

1 INTRODUCTION

Open-weight language models are released faster than their provenance can be audited. Model cards may omit the base checkpoint used for fine-tuning, training-data disclosures are often incomplete, and nominally independent models may share substantial portions of their corpora. Existing forensic methods address parts of this problem. Weight-space approaches recover fine-tuning trees when parameters are shared (Horwitz et al., 2025; Zeng et al., 2024; Tong et al., 2026; Wu et al., 2025), while representation-based methods such as REEF test whether a suspect model derives from a specific source model (Zhang et al., 2025b). These methods assume parameter inheritance. Behavioral phylogenies such as PhyloLM (Yax et al., 2025) apply to arbitrary model pairs, but yield uncalibrated structure: a dendrogram reveals which models cluster, not what level of similarity should be expected from a shared base, overlapping training data, or similar task competence.

We study a simple training-free fingerprint: the top-k vocabulary projections of late hidden states on 250 fixed knowledge probes, compared by Jaccard overlap of decoded token strings. We evaluate it on a 32-model pool with documented relationships spanning multiple levels of relatedness: Pythia models (Biderman et al., 2023), which share training data, order, and recipe but not weights; DeepSeek-R1 distillations (Guo et al., 2025), which inherit published base checkpoints; same-lineage-group models; successive generations from the same developer; and models with no documented relationship. An extended witness pool adds models trained on the Pile by three organizations, with different tokenizers (Cerebras-GPT (Dey et al., 2023)), codebases (GPT-NeoX-20B (Black et al., 2022)), and architecture classes (RWKV-4-Pile (Peng et al., 2023)).

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.

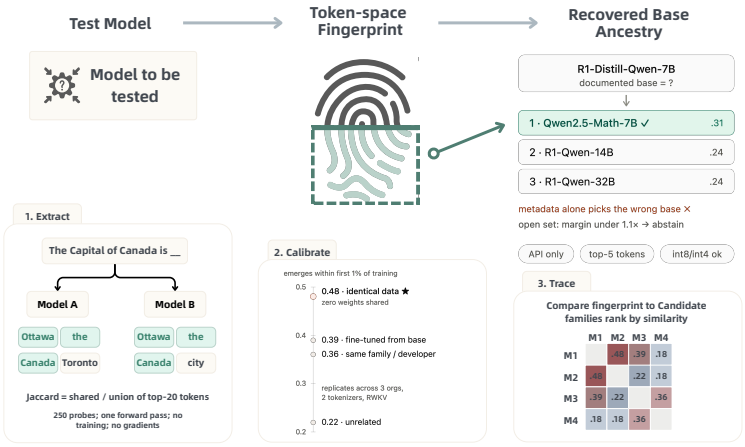

Figure 1: Overview: each model answers 250 fixed knowledge probes; top-k token sets at matched depths are compared by Jaccard overlap; pairwise scores are calibrated against documented model relationships.

Calibration yields three main results. First, fingerprint similarity follows documented model relatedness and remains informative across organizations, tokenizers, and architectures. Second, nearest-neighbor retrieval consistently narrows each R1 distillation to a small lineage neighborhood, while also exposing the difficulty of distinguishing an exact parent from closely related alternatives using output behavior alone. Third, the signal becomes stronger toward the output distribution and remains usable under restricted top-k access and model quantization. Together, these results characterize both the utility and the limitations of token-space fingerprints for model-provenance analysis.

Overall, this work introduces a training-free token-space fingerprint with an explicit shared-vocabulary null, evaluates it across documented model relationships, and studies its behavior across training stages, model depth, access constraints, and quantization. It further shows how the resulting similarity structure can support lineage retrieval while avoiding claims of exact provenance that output-space evidence cannot reliably establish.

2 Related Work

Weight- and representation-space fingerprints. HuRef derives human-readable invariants from parameter matrices (Zeng et al., 2024); Model Tree Heritage Recovery reconstructs fine-tuning trees from weights (Horwitz et al., 2025); SeedPrints traces lineage to random initialization (Tong et al., 2026); gradient-based features support lineage group classification (Wu et al., 2025); and REEF compares paired activations using CKA (Kornblith et al., 2019) to test derivation hypotheses (Zhang et al., 2025b). These methods require parameter access and primarily target settings with parameter inheritance. Our fingerprint requires only forward passes and also applies when models share no weights.

Behavioral phylogenies and output analysis. PhyloLM infers phylogenies from sampled-output similarity across a large model pool (Yax et al., 2025), while refusal directions have been proposed as behavioral fingerprints (Xu & Sheng, 2026). Our output-level fingerprint belongs to this broader class, but focuses on calibrating pairwise similarity against documented model relationships. We also compare against a probability-overlap baseline based on the PhyloLM distance under the same model pool and evaluation protocol. Layerwise logit trajectories have been used to detect benchmark contamination within individual models (He et al., 2025); we use related readouts for cross-model comparison.

Distillation and training-data detection. Membership and dataset-inference methods test whether particular examples or datasets were used in training (Shi et al., 2024; Zhang et al., 2025a), while distillation detectors test specific teacher–student hypotheses (Shi et al., 2026). Watermark persistence has also been studied as a deterrent to unauthorized distillation (Pan et al., 2025). These

methods typically address instance-level or pair-specific questions. Our method is complementary: it produces a global similarity structure from 250 forward passes per model, which can be used to generate lineage hypotheses for subsequent targeted analysis.

3 METHOD

3.1 FINGERPRINT EXTRACTION

Given a causal language model with L layers, unembedding matrix W_U, and final normalization $\mathrm{norm}(\cdot)$, we use each probe p as a raw continuation prompt, without a chat template, and record the last-position hidden state $h_\ell(p)$ at depths $\ell \in \{L/4, L/2, 3L/4, L-2\}$. We project each state through the model's output head, $z_\ell(p) = W_U \,\mathrm{norm}(h_\ell(p))$, following the logit-lens formulation (nostalgebraist, 2020; Belrose et al., 2023). For all 32 models, we also record the top-100 tokens from the final output logits.

We remove *glitch tokens* whose unembedding norm is above $5\times$ or below $0.2\times$ the vocabulary median, decode the remaining top-100 token ids, strip surrounding whitespace, and discard empty strings. We retain the top 100 tokens to support the k ablation, but use k=20 by default. This choice balances specificity against instability near the cutoff. Results vary by less than 0.04 in lineage group AUC over $k \in \{10, 20, 50, 100\}$ (Appendix C). We use unweighted token sets and leave rank-weighted variants to future work.

3.2 PROBE SUITE

The probe suite contains 250 prompts across 13 domains, including common and long-tail geography, science, history, code, mathematics, multi-step recall, culture, Chinese-language knowledge, multilingual prompts in nine languages, and contemporary facts. Per-domain counts and examples appear in Appendix B, and the full suite is included in the released code.

The domain taxonomy and probe set were fixed before computing any cross-model similarity and were not modified in response to the results. Common-knowledge probes provide a shared baseline, while long-tail and language-specific probes contribute more strongly to discrimination. Domain-level aggregation is used to identify which content areas contribute to similarity between a given pair of models.

3.3 SIMILARITY, NULL, AND STATISTICS

For models a and b at depth ℓ, we define similarity as the mean Jaccard overlap of their top-k decoded-string sets across probes:

$$S_\ell(a,b) = \frac{1}{|P|} \sum_{p \in P} \frac{|F_{a,\ell}(p) \cap F_{b,\ell}(p)|}{|F_{a,\ell}(p) \cup F_{b,\ell}(p)|} \tag{1}$$

When $L_a \neq L_b$, internal depths are matched by fractional position.

String-level comparison permits evaluation across different vocabularies, but shared tokenizers and globally frequent tokens can inflate similarity independently of probe content. We therefore estimate a pair-specific *mismatched-probe null*, $\tilde{S}(a,b)$, by pairing model a's response to probe i with model b's response to a different probe $j \neq i$, using 200 random mismatched pairs. We report both the raw score S and the *excess similarity* $(a,b) - \tilde{S}(a,b)$.

At near-final depth, the null averages 0.089 for same-tokenizer pairs and 0.054 for cross-tokenizer pairs; at $3L/4$, the corresponding values are 0.068 and 0.012. This correction removes probe-independent lexical overlap, but not differences caused by alternative tokenization of semantically equivalent strings. Cross-tokenizer similarity may therefore remain underestimated.

Statistical significance is assessed at the model level. We shuffle lineage-group labels over the 32 models for 10^4 permutations and recompute the within-group versus between-group similarity gap, preserving dependence among pairwise scores. Cohen's d is reported only as a descriptive pair-level effect size.

3.4 Model Pool

The calibration pool contains 32 open-weight models ranging from 0.6B to 32B parameters and covering nine developer families; the full list is provided in Appendix A. For analysis, we define 11 *lineage groups*, separating R1-Qwen, R1-Llama, and Qwen2.5-Math from their broader developer families because their documented training histories differ.

The pool includes Qwen3 (Qwen Team, 2025), Qwen2.5 (Qwen Team, 2024), Llama-3.1/3.2 (Llama Team, AI @ Meta, 2024), Mistral and Ministral (Jiang et al., 2023), Phi-3.5 and Phi-4 (Abdin et al., 2024b;a), InternLM2.5 (Cai et al., 2024), Gemma-2 (Gemma Team, Google DeepMind, 2024), Pythia (Biderman et al., 2023; Gao et al., 2020), and five DeepSeek-R1 distillations (Guo et al., 2025). It contains both base and post-trained models. The documented base checkpoints Qwen2.5-32B and Llama-3.1-8B are included in both the calibration and lineage-retrieval analyses.

Each experiment uses an explicitly defined pool. The *calibration pool* contains the 32 models used for all ladder and lineage group-level analyses. The *ancestry pool* adds Qwen2.5-Math-1.5B and Qwen2.5-14B as candidate bases, yielding 33 candidates for each query. The Pythia-deduped models and four additional base models—GPT-2-XL (Radford et al., 2019), OPT-6.7B (Zhang et al., 2022), OLMo-2-7B (OLMo et al., 2024), and Qwen2.5-7B—are used only as robustness controls and do not enter calibration statistics.

A separate *witness pool* contains GPT-NeoX-20B (Black et al., 2022), Cerebras-GPT 1.3–13B (Dey et al., 2023), Pythia-2.8B, and RWKV-4-Pile 3B/7B (Peng et al., 2023). These models provide same-corpus comparisons across organizations, tokenizers, and architecture classes and are used only in the witness analysis (Section 4.2).

4 Results

4.1 The Calibration Ladder

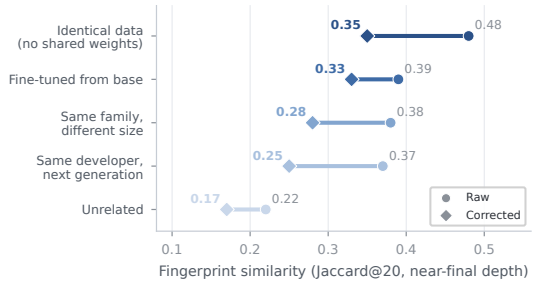

Figure 2: Mean fingerprint similarity across documented relationship categories, shown for raw scores (circles) and excess scores after vocabulary correction (diamonds). The highest-similarity category contains no shared parameters.

Figure 2 and Table 1 summarize the main calibration result. Mean similarity generally increases with documented model relatedness for both raw and excess scores. The two same-developer categories are close (difference $+0.014$, 95% CI $[+0.010, +0.019]$), so the results support three broader regimes: strongly related pairs, including identical-data and shared-base pairs (0.35 and 0.33 excess); same-developer pairs (0.25–0.28); and pairs with no documented relationship (0.17). Because the category distributions overlap, these values provide pool-level reference ranges rather than pair-level decision thresholds; individual retrieval results are evaluated using nearest-neighbor ranks and margins (§4.5).

The non-related category retains substantial similarity (0.22 raw; 0.17 excess), consistent with shared lexical and distributional structure across modern language models and motivating comparison against calibrated reference groups. The Pythia-deduped controls reproduce the identical-data result and isolate training order: cross-size similarity is nearly unchanged when the models use disjoint token streams and different orders (0.463 vs. 0.462). Within the resolution of this experiment, shared token order does not contribute detectably to the score (Appendix C).

	n	Raw	Excess
Relationship categories			
Identical data, no shared weights	3	0.48	0.35
Replication: deduped suite[†]	3	0.45	0.35
Same corpus, diff. order (cross-size)[†]	6	0.46	0.35
Same corpus, diff. order (size-matched)[†]	3	0.51	0.39
Fine-tuned from documented base	3	0.39	0.33
Same lineage group, different size	36	0.38	0.28
Same developer, next generation	30	0.37	0.25
No documented relationship	418	0.22	0.17
Lineage group structure			
ROC AUC (same vs. cross group)		0.859	0.847
Permutation p (model-level)		$< 10^{-4}$	$< 10^{-4}$
Lineage retrieval (R1 distills)	5	5/5 top-2; MRR 0.60	

Table 1: Results at near-final depth for the 32-model calibration pool. Probe-bootstrap 95% CIs lie within ± 0.013 of each category mean. The shared-base category contains the three distill↔base pairs for which the documented base is included in the pool: R1-Q-32B↔Q2.5-32B, R1-Q-7B↔Q2.5-Math-7B, and R1-L-8B↔L3.1-8B. The categories cover 490 of the $\binom{32}{2} = 496$ pairs. The remaining six pairs are among the four R1-Qwen distillations, which share a teacher but use different base models and are therefore excluded from the category analysis. [†]Control results from the Pythia-deduped suite, which is not included in the calibration pool; the corresponding 2×2 comparison isolates the effect of training order (Appendix C).

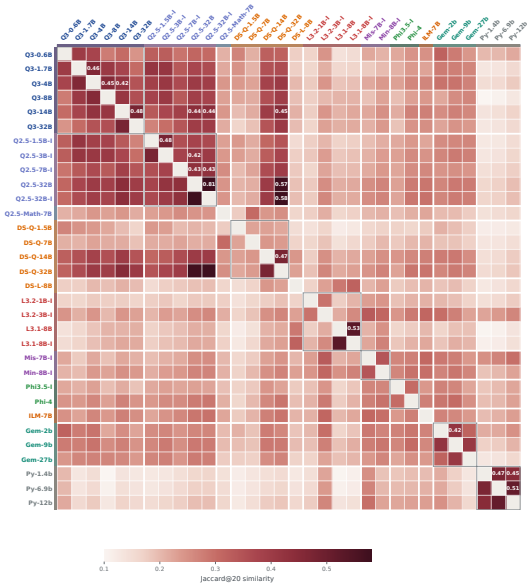

Figure 3: Raw pairwise fingerprint similarity (Jaccard@20) at near-final depth for the 32 calibration models, ordered by lineage group. Within-lineage group blocks and several fine-tune↔base pairs are visible. The strongest non-identical pair is Qwen2.5-32B base↔Instruct (0.81); other high-similarity pairs include R1-Q-32B↔Qwen2.5 models (0.57–0.58) and the identical-data Pythia pairs (up to 0.51).

4.2 External Witnesses: The Same-Data Signal Across Organizations, Tokenizers, and Architectures

The identical-data signal within Pythia could reflect shared tokenizer, training recipe, or token order rather than training data alone. We therefore evaluate three external witness groups that vary these factors (Table 2).

GPT-NeoX-20B (Black et al., 2022), trained on the Pile by the same organization as Pythia but in a separate project, achieves a mean excess similarity of 0.43 with Pythia at the output level, matching the Pythia-internal value. Cerebras-GPT (Dey et al., 2023) (1.3B–13B), trained on the Pile by a different organization with a different tokenizer and training recipe, reaches 0.35. This value is lower than the same-tokenizer witness results, consistent with reduced string overlap across tokenizers, but remains substantially above the reference value of 0.15 for pairs with no documented relationship. RWKV-4-Pile (Peng et al., 2023), a recurrent non-Transformer model trained on the Pile, reaches 0.43 despite differing in tokenizer and architecture class. All witness comparisons use the output distribution for consistency.

Witness group	Main difference	n	Excess
Pythia internal	baseline	3	0.43
NeoX-20B↔Pythia	project	3	0.43
RWKV↔Pythia	architecture class	6	0.43
Cerebras↔Pythia	tokenizer, organization, recipe	12	0.35
No documented relationship	—	6	0.15

Table 2: Output-level excess similarity across same-corpus witness groups. The signal remains above the reference level across projects, organizations, tokenizers, and architecture classes. The lower Cerebras value is consistent with reduced decoded-string overlap across tokenizers.

We further examine when the signal emerges during training (Figure 4). For 121 probes with unambiguous factual answers, we define *probe accuracy* as the fraction for which a gold answer string appears among the model's top-20 decoded tokens. This measure uses the same forward pass as the fingerprint but evaluates factual answer recovery rather than cross-model overlap.

Across publicly released Pythia checkpoints, excess similarity between the 1.4B and 6.9B models reaches 0.25 at step 1000 (0.7% of training), peaks at 0.43 by step 8000 (5.6%), and declines to 0.34 at the final checkpoint. Over the same trajectory, similarity to Qwen2.5-7B remains between 0.05 and 0.10. At step 1000, probe accuracy is 0.02 while excess similarity is already 0.25. This temporal separation suggests that the early similarity signal is not explained solely by convergence in factual competence.

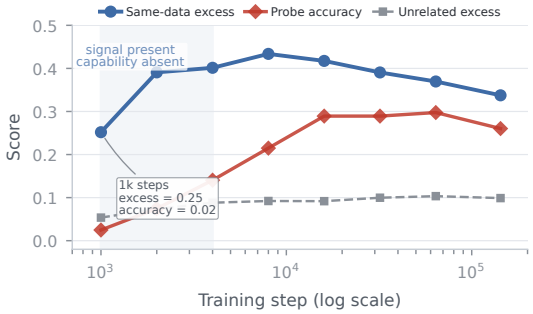

Figure 4: Training trajectory on log-scaled steps. Excess similarity between Pythia-1.4B and Pythia-6.9B rises within the first 1% of training, while probe accuracy remains near zero. Similarity to an external reference remains low.

4.3 Lineage Group Structure at the Model Level

Within-group similarity exceeds cross-group similarity for every lineage group containing multiple models. A model-level permutation test yields $p < 10^{-4}$ for both raw and excess scores, with observed gaps 10.2 and 9.5 standard deviations above the permutation mean, respectively. Pair-level Cohen's d is 1.81 and is reported descriptively.

Using similarity to distinguish same-group from cross-group pairs gives an AUC of 0.859 for raw scores and 0.847 for excess scores. The result is not driven by a single model block: leave-one-lineage-group-out AUC ranges from 0.820 to 0.903; removing the full Qwen constellation gives

Configuration	Group AUC
Full pool (32 models, raw)	0.859
Full pool (excess-corrected)	0.847
Leave-one-lineage-group-out (range)	0.820–0.903
Drop Qwen constellation	0.832
Drop Pythia	0.820
Leave-one-model-out jackknife (range)	0.844–0.883
$k = 10,\ 20$	0.856, 0.859
$k = 50,\ 100$	0.839, 0.820

Table 3: Same-group discrimination across score correction, model-pool composition, resampling, and the choice of k at near-final depth.

0.832; removing Pythia gives 0.820; and the leave-one-model-out jackknife ranges from 0.844 to 0.883 (Table 3).

4.4 Same-Pool Baseline: Probability Overlap

Following the similarity used by PhyloLM (Yax et al., 2025), we compute the Bhattacharyya overlap $\sum_t \sqrt{P_a(t \mid p) P_b(t \mid p)}$ between output-token distributions and average it across probes. This baseline is evaluated on the 26-model subpool for which full probability vectors were retained.

On this subpool, probability overlap achieves a same-group AUC of 0.765, compared with 0.942 for the top-k set fingerprint at the same output depth. In the single lineage-retrieval case whose documented base is available in this subpool, both methods rank the base first, with top-1 margins of $1.05\times$ and $1.09\times$, respectively. These results indicate that top-k token identity is more discriminative on this pool, although the lineage comparison is limited to one query.

4.5 Lineage Retrieval and Neighborhood Ambiguity

For each of the five R1 distillations with an available documented base, we rank the remaining 33 models by fingerprint similarity (Table 4). The documented base appears within the top two in all five cases, with mean rank 1.8, MRR 0.60, and Hits@2 of 1.00. It ranks first in one case, R1-Q-7B→Qwen2.5-Math-7B, with a margin of $1.27\times$. In the other four cases, the top-ranked candidate is a same-developer relative.

For R1-Q-32B and R1-L-8B, the Instruct variant associated with the same base checkpoint ranks above the base. For R1-Q-14B, another R1 distillation initialized from a different Qwen base ranks first. For R1-Q-1.5B, Qwen3-0.6B ranks first. These cases show that the fingerprint often identifies a local lineage neighborhood, but does not reliably distinguish the exact parent from closely related alternatives using output behavior alone.

The 7B query provides a more specific retrieval case. Its documented base is Qwen2.5-Math-7B rather than the general-purpose Qwen2.5-7B. A metadata baseline based on tokenizer lineage group and parameter count selects the wrong model, whereas the fingerprint ranks the math-specialized base first. This result suggests that behavioral similarity can distinguish some specialized and general-purpose variants that coarse metadata cannot.

Candidate-removal analysis. We consider two candidate-removal conditions. In *open-a*, the documented base is removed; the top-ranked candidate remains a same-developer relative in all five cases. In *open-b*, the base, co-distilled siblings, and same-base Instruct variants are removed; four of five top-1 margins fall below $1.01\times$, while the remaining margin is $1.09\times$. These results suggest that small nearest-neighbor margins may indicate the absence of a close candidate, but a calibrated abstention rule would require an independent validation set.

Query	Top-1	Top-2	Base rank	Relation
R1-Q-32B	Q2.5-32B-Inst (.58)	**Q2.5-32B** (.57)	2	same-base
R1-L-8B	L3.1-8B-Inst (.32)	**L3.1-8B** (.29)	2	same-base
R1-Q-7B	**Q2.5-Math-7B** (.31)	R1-Q-14B (.24)	1	exact
R1-Q-14B	R1-Q-32B (.47)	**Q2.5-14B** (.46)	2	co-distill
R1-Q-1.5B	Q3-0.6B (.27)	**Q2.5-M-1.5B** (.26)	2	next-gen

Table 4: Lineage retrieval over 33 candidates at near-final depth. Bold denotes the documented base. The base ranks within the top two for all five queries, but ranks first in only one case. Relations describe the top-ranked candidate relative to the documented base.

4.6 Depth Ablation and Operational Conditions

Figure 5 evaluates four internal readouts and the final output distribution across all 32 models. Same-group AUC is 0.72 at quarter depth, 0.70 at half depth, 0.77 at three-quarter depth, 0.86 at $L-2$, and 0.90 at the output. Pair-level Cohen's d follows the same overall pattern and reaches 2.22 at the output.

The final output distribution is therefore the strongest readout in this evaluation. We nevertheless use $L-2$ for the primary internal-state analyses because fractional depths are not necessarily aligned across architectures and scales: the depth at which intermediate predictions approach the final distribution varies across model configurations (Appendix C).

Truncation and temperature. At the output level, restricting the fingerprint to the top $k{=}5$ tokens retains a same-group AUC of 0.87. Because the fingerprint depends only on token ranking, it is invariant to strictly monotone transformations of the logits, including temperature scaling applied before top-k selection (Proposition 1, Appendix E).

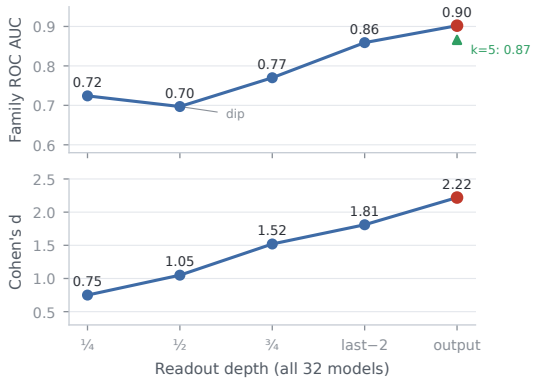

Figure 5: Same-group discrimination across readout depths for all 32 calibration models. Performance is highest at the final output distribution; output-level AUC remains 0.87 when restricted to the top 5 tokens.

Domain-level aggregation provides a profile of which probe categories contribute to pairwise similarity. For example, Qwen↔Qwen similarity exceeds Qwen↔Llama similarity by $6.2\times$ on Chinese-language probes, whereas the difference is smaller on code probes (Appendix C).

5 Robustness

Tokenizer stratification. In the public model ecosystem, same-lineage group pairs generally also share a tokenizer, so cross-tokenizer same-lineage group AUC cannot be estimated directly because no positive pairs exist. We therefore assess tokenizer effects through three complementary analyses. First, within the 16-model Qwen subpool, whose models share the same BPE vocabulary core, lineage-group discrimination remains above chance (AUC 0.73) when separating Qwen2.5, Qwen3, and R1-Qwen models. Second, subtracting the mismatched-probe null reduces the contribution of

probe-independent lexical overlap while retaining an AUC of 0.847. Third, the external witness analysis provides same-corpus comparisons across tokenizers: Cerebras↔Pythia pairs achieve an excess similarity of 0.35, compared with 0.15 for pairs with no documented relationship (Table 2). These results indicate that tokenizer sharing contributes to raw similarity but does not fully explain the observed structure.

Shared-vocabulary null. Table 1 reports both raw and excess scores. The correction lowers absolute similarity while preserving the main within-group structure and the lineage-retrieval results. The estimated null is strongly tokenizer-dependent, averaging approximately 0.03 for Llama pairs and 0.13 for Qwen2.5 and Pythia pairs. Raw scores are therefore not directly comparable across tokenizer groups. Excess scores improve comparability by removing probe-independent overlap, although they do not correct for semantically equivalent strings that are segmented differently across vocabularies (Appendix C).

Base-model controls. We test whether the results could arise from systematic differences between base and instruction-tuned models. First, after adding Qwen2.5-7B, its base↔Instruct similarity is 0.76, close to the Qwen2.5-32B base↔Instruct value of 0.81. By contrast, the cross-lineage group pair Qwen2.5-7B↔Llama-3.1-8B-Instruct scores 0.20, near the reference level for pairs with no documented relationship. This suggests that post-training preserves substantially more of the fingerprint than changing model lineage.

Second, three additional base models—GPT-2-XL, OPT-6.7B, and OLMo-2-7B—form 24 cross-suite base↔base pairs with mean similarity 0.23, close to the overall reference mean of 0.22 and well below the Pythia identical-data mean of 0.48. Thus, a generic response pattern associated with base models does not reproduce the identical-data signal. The highest control similarities are GPT-2↔OPT (0.40) and OPT↔Pythia (0.34–0.37), consistent with their documented corpus relationships (Zhang et al., 2022; Gao et al., 2020); Qwen↔Pythia base pairs score 0.14–0.18.

Stability. The main same-group discrimination result remains stable under leave-one-lineage-group-out evaluation, removal of the full Qwen or Pythia blocks, and a leave-one-model-out jackknife (Table 3); each block-removal permutation test gives $p < 10^{-3}$. Removing any single probe domain yields AUC values between 0.858 and 0.870, and the documented base remains within the top two for all five lineage queries. AUC changes by less than 0.04 across $k \in \{10, 20, 50, 100\}$. Independent fp32 and bf16 extraction pipelines agree at Jaccard 0.98 (Appendix C).

Capability-convergence control. We fit the following regression:

$$S_{ab}^{\text{excess}} = \beta_0 + \beta_1 \mathbb{I}[g_a = g_b] + \beta_2 \Delta C_{ab} + \beta_3 |\Delta \log P_{ab}| + \varepsilon_{ab}. \tag{2}$$

Using two-way cluster-robust standard errors at the model level, the same-group coefficient remains positive after controlling for capability and parameter scale ($\beta_1 = 0.103$, $z = 6.46$). Capability gap is also associated with similarity ($\beta_2 = -0.24$, $z = -4.86$), but capability gap alone yields an AUC of 0.49 for same-group discrimination. These results indicate that measured capability similarity does not explain the observed lineage-group structure.

6 Conclusion

Top-k token projections on a fixed probe set provide a lightweight fingerprint for comparing language models in output space. After calibration against documented relationships, the resulting similarity structure distinguishes broad levels of model relatedness, retrieves the documented base within the top two candidates for all five R1 distillations, and identifies specialized bases that coarse metadata can miss. The signal is strongest near the output distribution, remains effective with top-5 token access, and is stable under int8 and int4 quantization. It also persists across organizations, tokenizer families, and architecture classes, and emerges early in training. These results position token-space fingerprints as a practical tool for lineage retrieval and model-provenance analysis.

7 AI Disclosure

In this work, we used generative AI tools for polishing writing and coding. We have not used generative AI tools for generating data, and providing mathematical claims are not applicable to this

work. Additionally, we used generative AI tools for ¡tasks with recommended disclosure¿. We have reviewed all AI-assisted work. [Elaborate. For example, "we checked LLM-generated research ideas for potential plagiarism through a manual literature survey", "LLM-generated code was verified and tested for correctness by 2 authors", etc.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024a.

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024b.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022.

Zheng Cai et al. InternLM2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zirui He, Haiyan Zhao, Yingcong Li, Ali Payani, and Mengnan Du. LogitTrace: Detecting benchmark contamination via layerwise logit trajectories. *arXiv preprint arXiv:2509.20909*, 2025.

Eliahu Horwitz, Asaf Shul, and Yedid Hoshen. Unsupervised model tree heritage recovery. In *International Conference on Learning Representations*, volume 2025, pp. 47900–47918, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru`, 2020.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. Can llm watermarks robustly prevent unauthorized knowledge distillation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13228–13251, 2025.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Qin Shi, Amber Yijia Zheng, Qifan Song, and Raymond A Yeh. Knowledge distillation detection for open-weights models. *Advances in Neural Information Processing Systems*, 38:119758–119782, 2026.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Yao Tong, Haonan Wang, Siquan Li, Kenji Kawaguchi, and Tianyang Hu. SeedPrints: Fingerprints can even tell which seed your large language model was trained from. In *International Conference on Learning Representations (ICLR)*, 2026.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Gradient-based model fingerprinting for llm similarity detection and family classification. *arXiv preprint arXiv:2506.01631*, 2025.

Zhenyu Xu and Victor S. Sheng. A behavioral fingerprint for large language models: Provenance tracking via refusal vectors. *arXiv preprint arXiv:2602.09434*, 2026.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. Phylolm: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 64792–64844, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a2e28663712d5a3429a98918c3058f7b-Paper-Conference.pdf.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37:126332–126362, 2024.

Hengxiang Zhang, Hyeong Kyu Choi, Sharon Li, and Hongxin Wei. Detecting distillation data from reasoning models. *arXiv preprint arXiv:2510.04850*, 2025a.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 48092–48117, 2025b.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix

B Model Pool

The pool was constructed backward from the evidence requirements: every rung of the relatedness ladder needs pairs whose relationship is *documented*, not inferred. Pythia supplies the identical-data rung (public training order, no weight sharing across sizes); the R1 distillations supply the shared-base rung (bases named in the release report); Qwen2.5→Qwen3 supplies successive generations under one developer; and the remaining families supply same-lineage group and unrelated pairs. Sizes span 0.6B–32B so that scale is represented within families rather than confounded across them.

For analysis we track 11 *lineage groups* rather than 9 developer families: R1-Qwen, R1-Llama, and Qwen2.5-Math are split off because their documented ancestry differs from their developers' main lines—R1 distills descend from Qwen/Llama bases but were post-trained by DeepSeek, and the Math models were continued-pretrained on a specialized corpus. Treating these as members of their surface families would contaminate both the lineage group statistics (a distill is not a sibling) and the ancestry experiment (the query would share a label with its own answer).

Each model carries exactly one role (table below). *Calibration* models generate every lineage group and ladder statistic; *ancestry candidates* are documented bases added only so the ancestry queries have their answers available; *witness*, *base-control*, *dedup-control*, and *quantization-control* models never enter calibration statistics—this separation is what keeps the headline numbers stable as evidence accumulates. All models are evaluated with identical raw continuation prompts and no chat template; the single exception is the teacher-probe experiment (Appendix C), where the API constraint forces chat formatting and all local comparisons are re-extracted to match.

C Probe Suite

The suite balances two opposing requirements. *Floor* probes (common geography, basic science, canonical history) are answered identically by every competent model; they contribute little discrimination but anchor the score—a model that fails them is out of distribution for the method, and their stability separates rank noise from real disagreement. *Discriminative* probes (niche capitals, long-tail facts, Chinese-language knowledge, contemporary events) are where training distributions actually differ; ablations confirm these carry the lineage group signal (removing common-knowledge domains barely moves AUC, while the overall structure degrades if all long-tail domains are removed together). Code probes use raw multi-line continuation contexts rather than natural-language descriptions, so they probe code-corpus statistics rather than instruction-following. Multilingual probes (9 languages) and the 20 Chinese probes give the domain profile its resolution: they are the axis along which Qwen and Llama separate most sharply ($6.2\times$; Appendix C).

Two disciplines govern the suite. First, it was authored from the domain taxonomy *before any cross-model similarity was computed* and has not been modified since—no probe was added, removed, or reworded in response to results. Second, the 15 contemporary probes (e.g., current office-holders) date the fingerprint to a knowledge-cutoff window; comparisons across models with very different cutoffs should weight the leave-one-domain-out results (lineage group AUC 0.858–0.870 without any single domain, including this one), which show no conclusion depends on the temporal slice.

A 121-probe subset with unambiguous short factual answers (e.g., "The capital of France is" → Paris) doubles as the *gold key* for the capability score used in the trajectory analysis and confound regression (§4.2): capability is the fraction of gold-key probes whose answer string appears in the model's top-20 tokens. The key was authored from probe text alone, blind to model outputs.

The probe list follows; newline characters in code probes are rendered as \n, and three probes in scripts without font support here (two Arabic, one Russian) are shown as placeholders—the complete list ships with the code release.

Geography (common). *The capital of France is · The capital of Japan is · The capital of Brazil is · The capital of Australia is · The capital of Egypt is · The capital of India is · The capital of Canada is · The capital of South Korea is · The capital of Mexico is · The capital of Germany is · The largest country*

Model	Family	L	$\|V\|$	Variant
Cerebras-GPT-1.3B	Cerebras	24	50k	base / witness pool
Cerebras-GPT-2.7B	Cerebras	32	50k	base / witness pool
Cerebras-GPT-6.7B	Cerebras	32	50k	base / witness pool
Cerebras-GPT-13B	Cerebras	40	50k	base / witness pool
DeepSeek-R1-Distill-Llama-8B	DS-Llama	32	128k	post-trained / calibration
DeepSeek-R1-Distill-Qwen-1.5B	DS-Qwen	28	151k	post-trained / calibration
DeepSeek-R1-Distill-Qwen-7B	DS-Qwen	28	152k	post-trained / calibration
DeepSeek-R1-Distill-Qwen-14B	DS-Qwen	48	152k	post-trained / calibration
DeepSeek-R1-Distill-Qwen-32B-int4	DS-Qwen	64	152k	int4 / quant. control
DeepSeek-R1-Distill-Qwen-32B-int8	DS-Qwen	64	152k	int8 / quant. control
DeepSeek-R1-Distill-Qwen-32B	DS-Qwen	64	152k	post-trained / calibration
gpt2-xl	GPT2	48	50k	base / base control
gemma-2-2b-it	Gemma	26	256k	post-trained / calibration
gemma-2-9b-it	Gemma	42	256k	post-trained / calibration
gemma-2-27b-it	Gemma	46	256k	post-trained / calibration
internlm2_5-7b-chat	InternLM	32	92k	post-trained / calibration
Llama-3.2-1B-Instruct	Llama	16	128k	post-trained / calibration
Llama-3.2-3B-Instruct	Llama	28	128k	post-trained / calibration
Llama-3.1-8B-Instruct	Llama	32	128k	post-trained / calibration
Llama-3.1-8B	Llama	32	128k	base / calibration
Mistral-7B-Instruct-v0.3	Mistral	32	32k	post-trained / calibration
Ministral-8B-Instruct-2410	Mistral	36	131k	post-trained / calibration
gpt-neox-20b	NeoX	44	50k	base / witness pool
OLMo-2-1124-7B	OLMo	32	100k	base / base control
opt-6.7b	OPT	32	50k	base / base control
Phi-3.5-mini-instruct	Phi	32	32k	post-trained / calibration
phi-4	Phi	40	100k	post-trained / calibration
pythia-1.4b	Pythia	24	50k	base / calibration
pythia-2.8b	Pythia	32	50k	base / witness pool
pythia-6.9b	Pythia	32	50k	base / calibration
pythia-12b	Pythia	36	50k	base / calibration
pythia-1.4b-deduped	Pythia-dd	24	50k	base / dedup control
pythia-6.9b-deduped	Pythia-dd	32	50k	base / dedup control
pythia-12b-deduped	Pythia-dd	36	50k	base / dedup control
Qwen2.5-Math-1.5B	Q25-Math	28	151k	base (math) / ancestry cand.
Qwen2.5-Math-7B	Q25-Math	28	152k	base (math) / calibration
Qwen2.5-1.5B-Instruct	Qwen2.5	28	151k	post-trained / calibration
Qwen2.5-3B-Instruct	Qwen2.5	36	151k	post-trained / calibration
Qwen2.5-7B-Instruct	Qwen2.5	28	152k	post-trained / calibration
Qwen2.5-7B	Qwen2.5	28	152k	base / base control
Qwen2.5-14B	Qwen2.5	48	152k	base / ancestry cand.
Qwen2.5-32B-Instruct-int4	Qwen2.5	64	152k	int4 / quant. control
Qwen2.5-32B-Instruct	Qwen2.5	64	152k	post-trained / calibration
Qwen2.5-32B	Qwen2.5	64	152k	base / calibration
Qwen3-0.6B	Qwen3	28	151k	post-trained / calibration
Qwen3-1.7B	Qwen3	28	151k	post-trained / calibration
Qwen3-4B	Qwen3	36	151k	post-trained / calibration
Qwen3-8B	Qwen3	36	151k	post-trained / calibration
Qwen3-14B	Qwen3	40	151k	post-trained / calibration
Qwen3-32B	Qwen3	64	151k	post-trained / calibration
rwkv-4-3b-pile	RWKV	32	50k	base / witness pool (output only)
rwkv-4-7b-pile	RWKV	32	50k	base / witness pool (output only)

Table 5: All models with their roles. *calibration*: the 32-model pool behind every family/ladder statistic; *ancestry cand.*: documented-base additions used only as ancestry candidates; *witness pool*: same-data witnesses from independent organizations (§4.2), excluded from calibration statistics; *base control* / *dedup control*: Robustness-section controls; *quant. control*: quantized variants of calibration models, used only in the quantization-envelope analysis. RWKV witnesses are non-Transformer and contribute output-level fingerprints only. The black-box teacher DeepSeek-V3 (671B; Appendix C) is accessed via its serving API and hosts no local checkpoint. L: layers; $|V|$: vocabulary size.

Domain	n	Example probe
Geography (common)	20	*The capital of France is*
Geography (niche)	20	*The capital of Burkina Faso is*
Science (common)	18	*The chemical formula for water is*
Science (niche)	17	*The Chandrasekhar limit for white dwarf stars*
Code / Technical	30	*def fibonacci(n): if n <= 1: retur*
History (common)	13	*The first president of the United States was*
History (niche)	12	*The Treaty of Westphalia was signed in*
Math / Reasoning	25	*The square root of 144 is*
Culture / Language	20	*The word 'Schadenfreude' means*
Chinese-specific	20	中国最长的河流是
Multilingual	20	*La capitale de l'Allemagne est*
Temporal / Contemporary	15	*The CEO of OpenAI is*
Long-tail / Obscure	20	*The national animal of Scotland is*
Total	250	

Table 6: Probe suite composition. Probes were authored from this domain taxonomy before any cross-model similarity was computed and never modified afterward; the complete list ships with the code release.

by area is · The longest river in the world is · The highest mountain in the world is · The largest ocean on Earth is · The smallest continent by area is · The largest desert in the world is · The country with the largest population is · The deepest point in the ocean is called · The largest island in the world is · The longest wall ever built is ·

Geography (niche). *The capital of Burkina Faso is · The capital of Bhutan is · The capital of Moldova is · The capital of Suriname is · The capital of Liechtenstein is · The second largest city in Kazakhstan is · The deepest lake in the world is located in · The smallest country in Africa by area is · The highest capital city in the world is · The longest river in Europe is · The largest landlocked country is · The country with the most time zones is · The driest inhabited continent is · The largest freshwater lake by surface area is · The only country that borders both the Atlantic and Indian oceans is · The strait separating Europe from Asia is called · The second highest mountain in the world is · The country with the most UNESCO World Heritage Sites is · The largest volcanic island is · The capital of the Maldives is ·*

Science (common). *The chemical formula for water is · The chemical symbol for gold is · The number of elements in the periodic table is approximately · The speed of light in meters per second is approximately · The force of gravity on Earth is approximately · The boiling point of water at sea level in Celsius is · The number of chromosomes in a human cell is · The powerhouse of the cell is called the · The molecule that carries genetic information is · The largest organ in the human body is · The closest star to Earth is · The number of planets in our solar system is · The largest planet in our solar system is · The unit of electrical resistance is the · The pH of pure water at room temperature is · The process by which plants convert sunlight to energy is · The speed of sound in air at room temperature is approximately · The distance from the Earth to the Sun is approximately ·*

Science (niche). *The Chandrasekhar limit for white dwarf stars is approximately · The enzyme that unwinds DNA during replication is called · The Krebs cycle produces a net total of · The oxidation state of manganese in potassium permanganate is · The fine-structure constant alpha is approximately · The neurotransmitter primarily responsible for reward is · The IUPAC name for aspirin is · The Schwarzschild radius of the Sun is approximately · The number of ATP molecules produced by oxidative phosphorylation is approximately · The electron configuration of chromium is · The de Broglie wavelength of an electron at 100 eV is approximately · The enzyme responsible for adding nucleotides during DNA replication is · The Hubble constant is approximately · The critical temperature of high-Tc superconductor YBCO is approximately · The protein that transports oxygen in blood is · The crystal field splitting in an octahedral complex is denoted · The Roche limit for a fluid satellite is approximately ·*

Code / Technical. *def fibonacci(n): \n if n <= 1: \n return · def binary_search(arr, target): \n left, right = 0, len(arr) - 1 \n while · import torch \n model = torch.nn.Linear(· with open('data.json', 'r') as f: \n data = · class Node: \n def __init__(self, val): \n self.val = val \n self.next = · from transformers import AutoModelForCausalLM \n model = AutoModelForCausalLM.from_pretrained(· async def fetch_data(url): \n async with aiohttp.ClientSession() as · df = pd.DataFrame({'name': ['Alice', 'Bob'], 'age': [25, 30]}) \n df.groupby(· SELECT * FROM users WHERE · SELECT COUNT(*) FROM orders GROUP BY ·*

*CREATE TABLE employees (\n id INTEGER PRIMARY KEY, \n name · const fetchData = async (url) => { \n const response = await · document.addEventListener('DOMContentLoaded', () => { \n const · const arr = [3, 1, 4, 1, 5]; \n arr.sort((a, b) => · fn main() { \n let mut v: Vec<i32> = Vec::new(); \n v.push(· #include <iostream> \n int main() { \n std::cout << · #!/bin/bash \n for file in *.txt; do \n echo · The time complexity of binary search is · The time complexity of quicksort in the average case is · In Python, a decorator is defined using the · The difference between a stack and a queue is · A hash table resolves collisions using · The CAP theorem states that a distributed system cannot simultaneously guarantee · In object-oriented programming, polymorphism means · The SOLID principle 'S' stands for · A mutex differs from a semaphore in that · The purpose of a garbage collector is to · def divide(a, b): \n return a / b \n # potential issue: · def average(nums): \n return sum(nums) / len(nums) \n # potential issue: · int arr[10]; \n for(int i=0; i<=10; i++) arr[i] = ·*

History (common). *The first president of the United States was · World War II ended in the year · The Berlin Wall fell in · The French Revolution began in · The first man to walk on the Moon was · The year Christopher Columbus reached the Americas was · The Declaration of Independence was signed in · The Renaissance began in · The inventor of the telephone was · The first emperor of Rome was · The year the Titanic sank was · The ancient civilization that built the pyramids of Giza was · The printing press was invented by ·*

History (niche). *The Treaty of Westphalia was signed in · The last emperor of the Byzantine Empire was · The Taiping Rebellion was led by · The Congress of Vienna took place in · The Battle of Thermopylae was fought between · The Meiji Restoration began in · The Edict of Nantes was issued by · The founder of the Mongol Empire was · The War of the Roses was fought between the houses of · The first Shogun of the Tokugawa shogunate was · The ancient city of Carthage was located in modern-day · The Rosetta Stone was discovered in ·*

Math / Reasoning. *The square root of 144 is · The value of pi to five decimal places is · What is 7 times 8? The answer is · What is 13 squared? The answer is · The factorial of 6 is · The integral of e^x dx is · The derivative of sin(x) is · The derivative of ln(x) is · The determinant of a 2x2 matrix [[a,b],[c,d]] is · The eigenvalues of the identity matrix are · Euler's identity states that e^(i*pi) + 1 = · The Pythagorean theorem states that · The fundamental theorem of calculus connects · A prime number is a number that · The Fibonacci sequence starts with · The number of legs on the animal that spins webs is · The language spoken in the country where the Eiffel Tower is located is · The currency used in the land of the rising sun is · The color you get when you mix red and blue is · The number of sides on a shape called a hexagon is · If a train travels at 60 mph for 2 hours, it covers · The planet known as the Red Planet is · The metal with the chemical symbol Fe is · The organ that pumps blood through the body is the · The gas that plants absorb from the atmosphere is ·*

Culture / Language. *The word 'Schadenfreude' means · The Japanese word 'tsunami' literally means · The Arabic word 'inshallah' translates to · The Hindi word 'namaste' means · The Latin phrase 'carpe diem' means · The French word for 'butterfly' is · The author of 'One Hundred Years of Solitude' is · The protagonist of 'Crime and Punishment' is · The author of 'The Tale of Genji' is · The author of 'Don Quixote' is · In Greek mythology, the god of the underworld is · In Norse mythology, the world tree is called · In Hindu mythology, the god of destruction is · The composer of 'The Four Seasons' is · The number of symphonies composed by Beethoven is · The painter of 'Starry Night' is · The sculptor of 'David' in Florence is · The Japanese dish made of vinegared rice and raw fish is called · The country where cricket originated is · The holy book of Islam is called ·*

Chinese-specific. 中国最长的河流是 · 中国面积最大的省份是 · 长城的东端起点是 · 中国五岳中最高的山是 · 秦始皇统一六国的年份是 · 唐朝的都城是 · 清朝的最后一位皇帝是 · 四大发明包括造纸术、火药、印刷术和 · 郑和下西洋开始于 · 红楼梦的作者是 · 三国演义中的三国是魏、蜀和 · 水浒传中一共有多少位好汉 · 李白最著名的诗之一是 · 中国第一颗人造卫星的名字是 · 青蒿素的发现者是 · 中国传统节日中秋节吃的食物是 · 中国象棋中将帅不能 · 京剧中的四大行当是生旦净 · 太极拳的基本理念来自 · 台湾海峡连接的两个海域是 ·

Multilingual. *La capitale de l'Allemagne est · Le plus grand fleuve de France est · L'auteur de 'Les Misérables' est · La tour Eiffel a été construite en · Die Hauptstadt von Österreich ist · Der höchste Berg Deutschlands ist · Die Formel für die kinetische Energie ist · Der Komponist der 'Mondscheinsonate' ist ·* 日本で一番高い山は · 日本の首都は · 源氏物の作者は · 日本で一番い川は · *La capital de Argentina es · El autor de 'Cien años de soledad' es ·* 대한민국의 수도는 · 한국에서 가장 높은 산은 · *A capital do Brasil é ·*

Temporal / Contemporary. *The CEO of OpenAI is · The programming language Rust was created by · The transformer architecture was introduced in the paper titled · Bitcoin was created by · The company that*

created ChatGPT is · The founder of Tesla is · The population of Earth in 2023 was approximately · The tallest building in the world as of 2024 is · The latest major version of Python is · The host city of the 2024 Summer Olympics was · The current Secretary-General of the United Nations is · The most recent country to join the European Union is · The AI model GPT-4 was released by · The company that developed the BERT model is · The deep learning framework PyTorch was developed by ·

Long-tail / Obscure. *The national animal of Scotland is · The only letter not appearing in any US state name is · The shortest war in history lasted approximately · The country with the most official languages is · The element with the highest melting point is · The first computer programmer is generally considered to be · The blood type known as the universal donor is · The language with the most native speakers in the world is · The phobia of long words is called · The only planet that rotates clockwise is · The inventor of the World Wide Web is · The hardest natural substance on Earth is · The tallest animal on Earth is · The only mammal capable of true flight is · The currency of Switzerland is · The SI unit of luminous intensity is · The Greek letter used to represent the golden ratio is · The number of bones in the adult human body is · The year the Internet was first made available to the public is · The astronomical unit is defined as the distance from ·*

D Extended Robustness Results

What the Correction Does to the Matrix

Comparing the raw matrix (Figure 3) with the corrected one (Figure 6) shows that the correction is a per-pair *recentering*, not a compression: every lineage group block, the fine-tune diagonal pairs, and the Pythia island survive, and every multi-member lineage group's intra-lineage-group mean stays above the corrected unrelated baseline of 0.166 (Table 7). What changes is comparability. The null component varies by a factor of five across families—0.027 for Llama pairs against 0.128 for Qwen2.5 and Pythia pairs—because it is driven by how much probability-mass-agnostic string overlap two vocabularies produce, which depends on vocabulary size, shared BPE merges, and each lineage group's propensity for generic high-frequency continuations. Raw values therefore systematically overstate the coherence of large-shared-vocabulary families relative to families with small or idiosyncratic token inventories. The correction costs a small amount of global discrimination (AUC 0.859→0.847, because cross-tokenizer unrelated pairs are easy negatives in raw scores) and buys interpretability: corrected values mean the same thing on both sides of a tokenizer boundary, which is the property the ladder and the witness comparisons rely on.

lineage group	n	Raw	Null	Excess
Pythia	3	0.476	0.127	0.349
Qwen2.5	10	0.446	0.128	0.318
Gemma-2	3	0.383	0.110	0.274
Qwen3	15	0.374	0.114	0.260
Mistral	1	0.343	0.062	0.280
Phi	1	0.303	0.071	0.232
DS-Qwen	6	0.271	0.065	0.206
Llama	6	0.280	0.027	0.253

Table 7: Per-lineage-group decomposition of intra-lineage-group similarity into null and excess components (near-final depth; n = intra-lineage-group pairs). The null spans 0.027–0.128 across families, so raw values are not comparable across tokenizer groups; excess values are. All families remain above the corrected unrelated baseline (0.166).

Training-order 2×2 (Pythia-deduped suite). The deduped suite completes a 2×2 over token order and size matching: cross-size pairs score 0.463 with identical token streams and 0.462 with disjoint streams and orders—a null order effect isolated with size held constant—while matched-size cross-suite pairs reach 0.51, an elevation attributable to scale matching (consistent with the convergence-depth effect below) rather than shared order. The order conclusion is immune to possible seed sharing because its cells compare across sizes, where initializations necessarily differ.

Prediction-convergence depth. Comparing each model's intermediate top-20 with its own final top-20 measures where along the depth axis the eventual prediction takes shape (Figure 7). In Qwen3, self-overlap at $3L/4$ falls with scale (0.39 at 0.6B to 0.01 at 32B); Gemma-2 replicates the

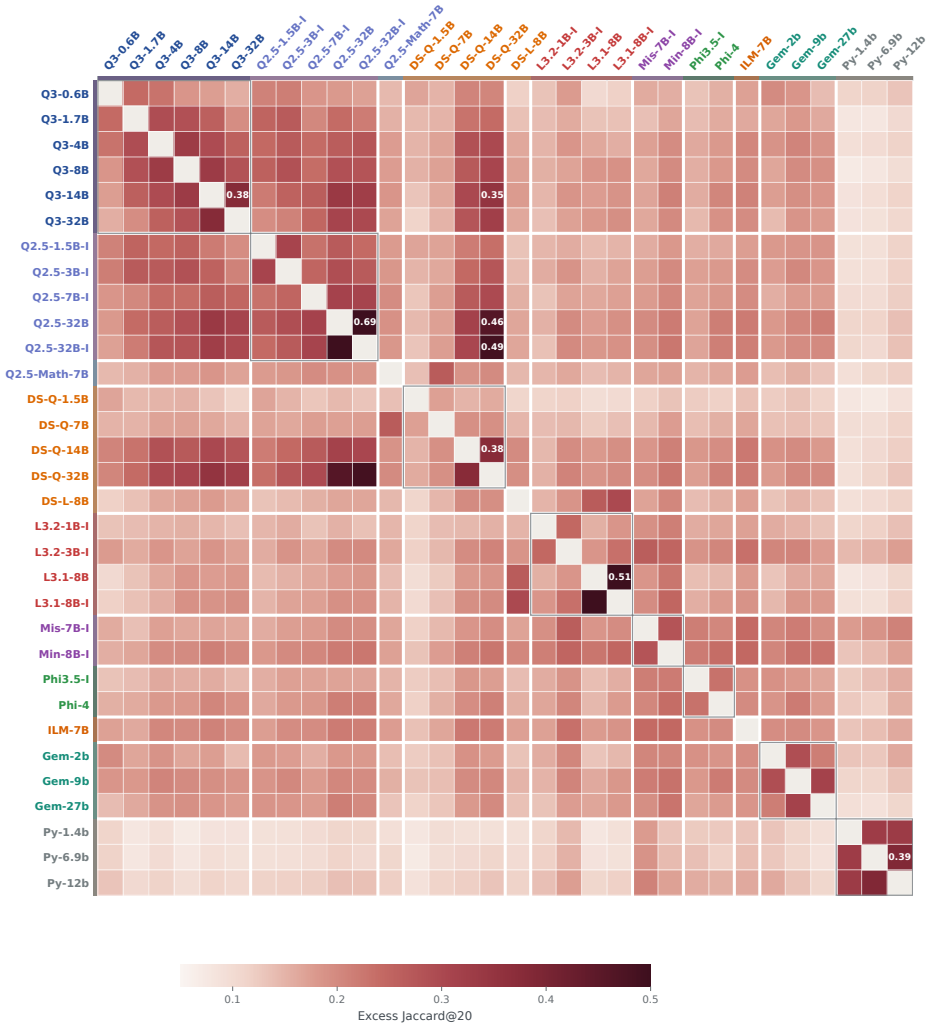

Figure 6: Excess-corrected similarity matrix (matched minus mismatched-probe null, near-final depth). Lineage group blocks, the fine-tune pairs, and the Pythia island persist after the shared-vocabulary correction; what the correction changes is the cross-lineage-group comparability of the values (Table 7).

trend (0.23 at 9B, 0.12 at 27B). The Pythia suite provides the control: across 1.4B, 6.9B, and 12B trained on identical data in identical order, self-overlap at $3L/4$ is flat (0.45, 0.47, 0.44). Whatever shifts convergence later co-varies with how larger models are trained, not merely with their size—and this misalignment of fractional depth across scales is why all cross-model results in the paper use near-final depth.

Domain profiles. On Chinese-language probes, Qwen↔Qwen similarity exceeds Qwen↔Llama by $6.2\times$ (probe-bootstrap 95% CI 4.7–$8.4\times$, at $3L/4$ over four Qwen and two Llama models); on code probes the same comparison largely closes (0.14 vs. 0.08).

Corrected ladder. Excess-corrected category means: identical data 0.349, fine-tuned from base (n=3) 0.332, same lineage group 0.276, successive generation 0.248, unrelated 0.166. The mismatched-probe null averages 0.090 for same-tokenizer pairs and 0.052 across tokenizers at near-final depth.

Corrected ancestry identification. Under excess similarity the outcome matches the raw analysis: the documented base is top-1 for one of five R1 distillations and top-2 for all five. The lineage-neighborhood ambiguity persists under correction: in the four non-top-1 cases, the displacing can-

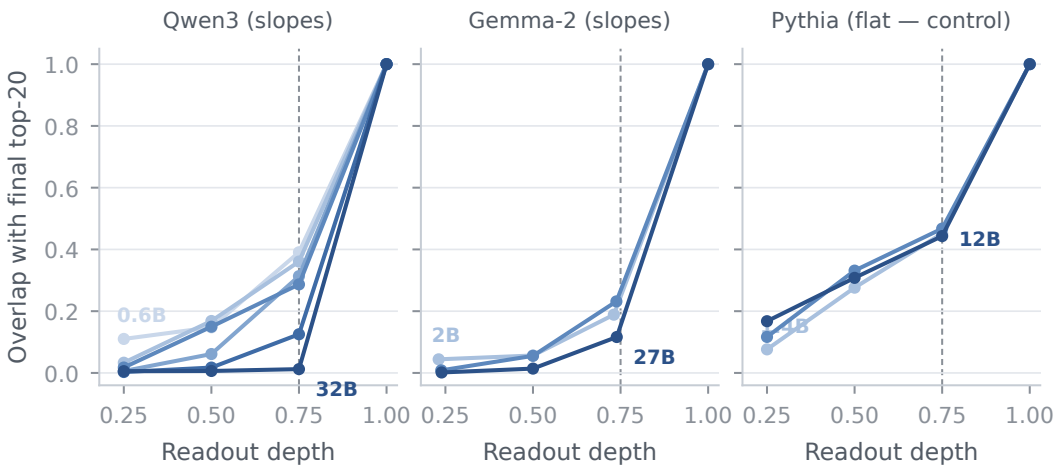

Figure 7: Self-overlap between each depth's top-20 and the model's own final top-20. Within Qwen3 and Gemma-2, convergence shifts later with scale; within Pythia (identical data, order, and recipe across sizes) it does not shift, implicating training configuration rather than parameter count alone.

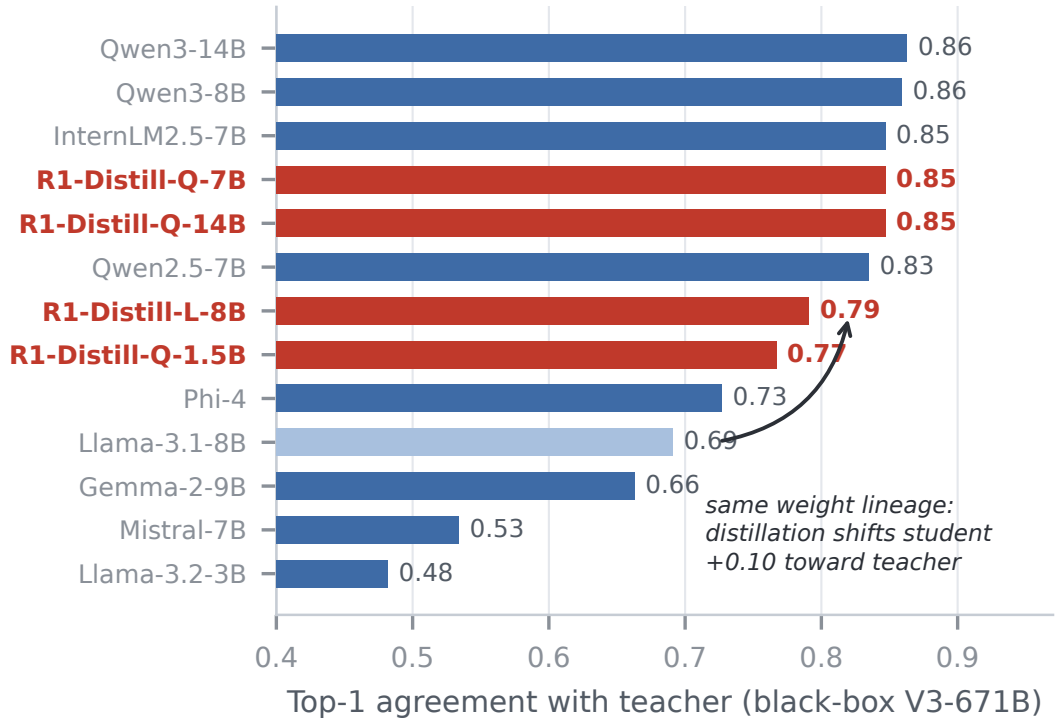

Figure 8: Black-box teacher probe: top-1 agreement of 13 chat-aligned local models with the sampled first-token behavior of DeepSeek-V3 (671B). Red: R1-distilled students. Group-level agreement is dominated by a shared-lab-data cluster, but the backbone-controlled pair (arrow) shows distillation displacing the Llama-based student $+0.10$ toward the teacher under zero weight sharing.

didate remains the same (same-base sibling, co-distillation, or next-generation model), with similar margins.

Query	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
R1-Qwen-1.5B	Q3-0.6B .27	**Q2.5-M-1.5B** .26	Q3-1.7B .25	Q2.5-1.5B .24	Q3-4B .24
R1-Qwen-7B	**Q2.5-M-7B** .31	R1-Qwen-14B .24	R1-Qwen-32B .24	R1-Qwen-1.5B .23	Q3-4B .23
R1-Qwen-14B	R1-Qwen-32B .47	**Q2.5-14B** .46	Q2.5-32B-I .40	Q3-14B .40	Q3-4B .38
R1-Qwen-32B	Q2.5-32B-I .58	**Q2.5-32B** .57	R1-Qwen-14B .47	Q2.5-14B .45	Q3-14B .45
R1-Llama-8B	L3.1-8B-I .32	**L3.1-8B** .29	R1-Qwen-14B .24	L3.2-3B .23	Min-8B .23

Table 8: Full top-5 candidate rankings for lineage retrieval over 33 candidates (raw, near-final depth). Bold: documented base. When the base is not rank 1, the displacing candidate is a same-developer relative: same-base Instruct sibling (marked "-I"), co-distillation, or next-generation model. The sole cross-developer entrant is Ministral-8B in the R1-Llama-8B list (rank 5, at the unrelated-baseline level 0.23).

Model-level significance. The permutation test shuffles lineage group labels over models, preserving all pairwise dependence under the null: raw $p < 10^{-4}$ (observed gap 10.2σ above the null), corrected $p < 10^{-4}$ (9.5σ). Dropping the Qwen constellation or Pythia leaves $p < 10^{-3}$ at $5{,}000$ permutations.

Glitch filtering. The unembedding-norm filter removes 0% of the vocabulary for most models and at most 1.6% (Qwen3-14B). Results are insensitive to the threshold because filtered tokens essentially never enter a top-100 list on our probes.

Teacher-side probe (black-box). DeepSeek-V3 was fingerprinted through its serving API by sampling 24 first tokens per probe at temperature 1 (no log-probabilities are exposed for this model lineage group). Because the API applies a chat template, all 13 local comparison models were re-extracted under their own chat templates, with thinking-mode blocks explicitly closed so the measured token is the answer token for every model. Top-1 agreement with the teacher: Qwen3-14B/8B 0.86, InternLM 0.85, R1-Distill-Qwen-7B/14B 0.85, Qwen2.5-7B 0.83, R1-Distill-Llama-8B 0.79, R1-Distill-Qwen-1.5B 0.77, Phi-4 0.73, Llama-3.1-8B 0.69, Gemma-9B 0.66, Mistral-7B 0.53, Llama-3.2-3B 0.48. The distilled-student group does not separate from the Chinese-lab data cluster as a whole (model-level permutation $p = 0.13$, $n = 13$); the backbone-controlled contrast (R1-Distill-Llama-8B vs. its base Llama-3.1-8B) shows the $+0.10$ displacement reported in the Discussion.

Prediction-convergence values. Self-overlap at $3L/4$ with the final top-20: Qwen3 0.6B/1.7B/4B/8B/14B/32B = 0.39/0.36/0.31/0.29/0.13/0.01; Gemma-2 2B/9B/27B = 0.19/0.23/0.12; Pythia 1.4B/6.9B/12B = 0.45/0.47/0.44.

E Formal Properties of the Score

Fix a probe set $\mathcal{P}$ and $k \in \mathbb{N}$. For a model m with vocabulary V_m, let $z_m(p) \in \mathbb{R}^{|V_m|}$ denote its logits (at a given readout) on probe p, and let $F_m(p) \subset \Sigma^*$ be the set of decoded, whitespace-stripped strings of the top-k entries of $z_m(p)$. The score is

$$S(a,b) \;=\; \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} J\big(F_a(p),\, F_b(p)\big), \qquad J(A,B) = \frac{|A\cap B|}{|A\cup B|}.$$

Proposition 1 (invariance under monotone score transformations). *Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing and let $z'_m(p) = g\big(z_m(p)\big)$ entrywise. Then $F'_m(p) = F_m(p)$ for every p, and hence S is unchanged.*

Proof. A strictly increasing g preserves the ordering of logit entries, so the top-k index set—and therefore the decoded string set—is identical. □

Temperature scaling (z/τ, $\tau > 0$), additive shifts, and softmax renormalization are all monotone, so the fingerprint is exactly invariant to them. This holds for any access mode that returns token *identities* in rank order (full logits, ranked log-probabilities, top-k lists); it does not extend to sampled tokens, whose distribution does change with temperature.

The mismatched-probe null. The null $\tilde{S}(a, b)$ is the same statistic computed on probe pairs (p_i, p_j), $i \neq j$, drawn uniformly at random. Its interpretation rests on one assumption: under the hypothesis that a model's top-k set does not depend on probe content, matched and mismatched pairings are exchangeable, so $\mathbb{E}[S] = \mathbb{E}[\tilde{S}]$ and the excess $S - \tilde{S}$ estimates the component of similarity that is *locked to probe content*. The null is far from the value implied by uniform random sets: two independent uniform k-subsets of a vocabulary of size $|V|$ have $\mathbb{E}|A \cap B| = k^2/|V|$, giving expected Jaccard $\approx 6.7 \times 10^{-5}$ for k=20, $|V|$=150k (and $\approx 2 \times 10^{-4}$ for $|V|$=50k)—three orders of magnitude below the observed nulls of 0.054–0.089. The null is therefore driven almost entirely by the frequency concentration of natural-language token distributions, not by vocabulary size, which is why it must be estimated empirically per pair rather than derived analytically, and why it varies by a factor of four across families (Table 7).

Proposition 2 (mechanism non-identifiability of behavioral scores). *Let $\Phi(m) = \{P_m(\cdot \mid p)\}_{p \in \mathcal{P}}$ be the map from a model to its probe-conditional output distributions, and let S be any score of the form $S(a, b) = f\big(\Phi(a), \Phi(b)\big)$. If two generative histories—(i) b inherits weights from a; (ii) a and b are trained independently on overlapping corpora—can produce the same pair $\big(\Phi(a), \Phi(b)\big)$, then no such S distinguishes them.*

Proof. Immediate: S factors through Φ, and Φ does not encode parameter history. □

The proposition is elementary, but it delimits what *any* behavioral score—ours, PhyloLM's, or a future one—can certify: behavioral evidence identifies training-distribution overlap, not the mechanism that produced it. Our pool contains an empirical witness that the antecedent is realized in practice: GPT-2↔OPT (independent training, overlapping web corpora) scores 0.40, the same value as the weight-inheriting fine-tune rung. This is why the paper's inferential unit is never a single pairwise score but structure around it—nearest-neighbor margins against a candidate pool, domain profiles, and the calibrated ladder—and why mechanism claims additionally require weight-space evidence where weights are available.

F Reproducibility

All experiments run on a single 80GB GPU or smaller. Extraction is one forward pass per probe per model; models up to 14B ran locally in fp16/fp32; the 27–32B models and the two exact base checkpoints (Qwen2.5-32B, Llama-3.1-8B) ran in bf16 on rented A100 instances. All 32 models thus have both internal-depth fingerprints and output-level top-100 fingerprints; the distinction matters only for the Bhattacharyya baseline (§4.4), which requires full probability vectors and is therefore restricted to the 26 models whose complete logit tensors were saved locally. A paired extraction of one model in fp32 and bf16 through two independent code paths agrees at Jaccard 0.98. Total extraction compute for the pool is under four GPU-hours (including the rental sessions). Probes, extraction and analysis code, per-model fingerprints, and all similarity matrices are released with the paper.