SLICEGRAPH: MAPPING PROCESS ISOMERS IN MULTI-RUN CHAIN-OF-THOUGHT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-run chain-of-thought (CoT) reasoning is usually summarized by final-answer votes, which cannot tell whether traces that agree on an answer repeat one internal route or reach it through different routes. We introduce **SliceGraph**, a post-hoc measurement object that makes this distinction measurable from completed CoT trajectories and their sparse internal activations: for one problem and one model, it links activation-similar trajectory segments in a mutual-kNN graph, reads biconnected graph regions as route blocks, and groups runs into *structural route families* by the blocks they visit. Two correct runs with the same normalized answer but different families are *structural process isomers* under the stated configuration. Across 954 problem–model cells from three 4B/8B models and five math and science benchmarks, 85.5% of cells contain multiple correct-only route families, 92.3% of eligible cells contain at least two families with two or more members, and 76.6% of correct-run pairs are cross-family under the reference configuration. We report this rate with its configuration envelope: at a fixed graph scale, 52.6% of pairs stay separated in at least 80% of threshold, weighting, and seed settings, whereas changing the scale itself reorganizes membership. Blinded annotation supports biconnected components as shared reasoning states and within-family strategy coherence, while cross-family boundaries only partially coincide with different human-named strategies. A label-free *route-collapsed consensus* flag—a majority answer supported by a single family—marks 7.8% of eight-run cohorts, whose accuracy is 33.6% versus 69.2% for the rest, whereas reweighting votes by family does not improve self-consistency. SliceGraph thus makes the structural support behind answer agreement a calibrated, auditable quantity with explicit limits on semantic interpretation, configuration dependence, and online use.

1 INTRODUCTION

Test-time sampling routinely draws many CoT trajectories per problem (Wang et al., 2023; Yao et al., 2023; Besta et al., 2024; Snell et al., 2025), and self-consistency aggregates them by counting final answers. Answer counts, however, do not describe how much the sampled trajectories overlap in their intermediate computation: several traces that support the same answer may revisit similar regions of activation space, or may follow different intermediate routes. Independently sampled runs can overlap in this way, and final-answer votes cannot distinguish repeated sampling of one route from agreement supported by several routes (Figure 1). Most internal analyses so far remain within a single run, studying semantic convergence, step-specific geometry, or late correctness signals (Wei et al., 2025; Sun et al., 2026), rather than the joint structure of many runs on one problem.

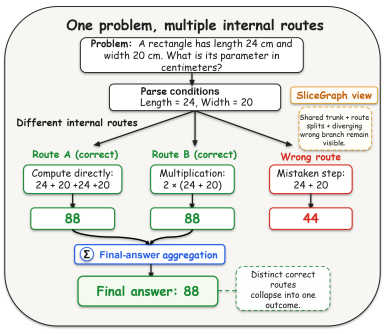


Figure 1: Answer votes cannot tell whether same-answer runs repeat one internal route or follow distinct route families.

We make this distinction measurable with **SliceGraph**. For one problem and one model (a *problem–model cell*) we sample $N{=}64$ CoTs and split each trace into 256-token slices. Each slice is represented by a sparse set of highly active internal units, and slices with similar activation sets are connected in a mutual-kNN graph. Recurring biconnected graph

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.

regions become *route blocks*; each completed CoT is summarized by the blocks it visits, and CoTs with similar visitation patterns are grouped into *structural route families*. Two correct CoTs that produce the same normalized answer but belong to different families are *structural process isomers*. Both terms are operational: a family boundary marks a difference in graph-measured route footprint under a stated configuration, and we reserve "different strategy" for pairs that human annotators judge to use different high-level strategies. SliceGraph is a measurement object built after sampling, not a decoding program.

Slice-level similarity is necessary because whole-run comparison is too coarse: two runs can share a long trunk and diverge at one fork, or reach the same answer by different routes. Prior graph and topology analyses of reasoning traces (Xiong et al., 2025; Zhou et al., 2026; Tan et al., 2025) operate on projected or text-embedded representations; inside our fixed scaffold, PCA-compressed and sentence-embedded slice representations compress or collapse the family geometry, whereas unprojected sparse activation-key Jaccard similarity retains it (Section 4.3).

The substantive question is then: when many CoTs solve the same problem, does correct-answer agreement repeatedly sample one internal route, or does it span several structurally distinct routes? Across 954 problem–model cells, 85.5% contain more than one correct-only route family, 92.3% of eligible cells contain at least two families with two or more members, and 76.6% of correct-run pairs are cross-family under the reference configuration. We treat this rate as a calibrated measurement rather than a universal constant: at a fixed graph scale, 52.6% of correct-run pairs remain separated in at least 80% of threshold, weighting, and seed settings, while changing slice granularity or neighbourhood size preserves aggregate multiplicity but reorganizes exact membership. Two blinded annotation studies validate the structural units before any geometric claim is made, and matched null controls, held-out atlases, and three replication models bound what the measurement can support.

Finally, we ask what the measurement is for. Reweighting answer votes by family does not improve self-consistency, but a label-free flag for *route-collapsed consensus*—a majority answer whose supporting runs all fall in one family—identifies a small subset of cohorts with substantially lower accuracy, giving a retrospective audit of how narrowly an apparent consensus is supported. Label-seeded reward overlays and typed-state dynamics are retained as secondary descriptive layers (Section 4.6).

Contributions.

- **A measurement object for multi-run reasoning.** SliceGraph builds a per-cell graph directly from sparse activation-key overlap between CoT slices and reads route blocks and structural route families from it. Graph construction never uses correctness labels, and we separate the correctness-conditioned diversity analysis from the label-free audit protocol.
- **Calibrated evidence of hidden route diversity.** Same-answer correct CoTs are usually structural process isomers. We report the headline rate together with its configuration envelope, non-singleton, split-half, and held-out checks, annotation-based semantic bounds, and cross-model replications.
- **A route-collapse audit and a negative result.** Route structure exposes structurally narrow consensus with low reliability, whereas family-based vote reweighting does not beat answer frequency; we state the retrospective boundary of this use.

2 Related Work

Answer aggregation and search programs versus measurement objects. Self-consistency (Wang et al., 2023) aggregates sampled final answers by frequency, while Tree-of-Thoughts (Yao et al., 2023) and Graph-of-Thoughts (Besta et al., 2024) organize inference through explicit search structures built before or during sampling. SliceGraph is instead a *measurement* object: it is built after sampling from local activation similarity across many completed runs, and it characterises process geometry rather than driving decoding.

Trajectory analyses. A line of work analyses LLM reasoning as structured motion in representation space, including semantic convergence from truncated trajectories (Wei et al., 2025), step-specific geometry (Sun et al., 2026), layer-wise displacement of the truth signal (Damirchi et al., 2026), hidden-state DAG probes (Zhong et al., 2026), and decision pivots shared by correct reasoning

paths (Cho et al., 2025). These works study individual trajectories or shared checkpoints across paths; we study the local activation-overlap graph of a whole sampled ensemble for one problem and the route footprints it induces.

Graph and topology pipelines. Prior process analyses use different representations and structural objects: Mapping the Minds (Xiong et al., 2025) clusters CoT paragraphs into NLI dependency graphs, Landscape of Thoughts (Zhou et al., 2026) represents reasoning states by distances to answer choices and projects them with t-SNE, and The Shape of Reasoning (Tan et al., 2025) runs persistent homology on step embeddings. Activation-set similarity has been used as a best-of-N selector in Neuron Agreement Decoding (Chen et al., 2025), and sparse activation caches have been used for label-free explore–exploit scoring of CoTs (Chen et al., 2026). SliceGraph builds its graph directly on unprojected sparse activation-key Jaccard similarity and analyses its combinatorial topology. Our representation ablation (Section 4.3) substitutes projected and text-embedded representations inside the SliceGraph scaffold; it is not an end-to-end benchmark of these pipelines.

3 SLICEGRAPH

SliceGraph has three scaffold layers and two secondary overlays. Section 3.1 builds a graph over CoT slices from sparse activation keys; Section 3.2 decomposes it into biconnected route blocks and assigns interpretive roles; Section 3.3 groups runs into structural route families and defines structural process isomers. Graph topology, block membership, and route families never use correctness labels in their construction. A label-seeded reward field and a typed-state transition model (Section 3.4) are descriptive overlays added on top of the fixed scaffold; they do not define families or isomers. All analyses operate on non-trivial blocks only. A glossary separating standard graph terms from the terms introduced here is given in Appendix A.

Two analysis protocols. We distinguish the graph scaffold from the population on which families are fitted. For the *structural-isomer analysis* (Section 4.4) the run-level co-visitation graph is restricted to normalized-correct runs and partitioned into families F_{correct}; this is a correctness-conditioned descriptive analysis of the diversity behind a shared correct answer. For *consensus auditing* (Section 4.5) a separate partition F_{all} is fitted to all sampled runs without correctness labels, which are used only to evaluate the audit.

3.1 SLICEGRAPH CONSTRUCTION

For each problem-model cell we sample $N{=}64$ separately sampled CoT runs from a precomputed sparse activation cache. Each *neuron* is a (layer, unit) pair indexed by a 32-bit key, and the active set at a given token is the top-ranked MLP neurons by the positive part of the post-gating intermediate (Appendix B.2). We first aggregate activations into 32-token rows, then aggregate every consecutive `sep_up` rows into one coarser reasoning slice; at the canonical `sep_up`$=8$, a graph node therefore summarises a 256-token segment. Magnitudes are discarded after retaining the most-active keys, so each analysed slice is a binary key set $\mathcal{K}_{r,t}$ with no projection or text-embedding step.

We connect two slices by their Jaccard overlap, $d_J(i,j) = 1 - |\mathcal{K}_i \cap \mathcal{K}_j|/|\mathcal{K}_i \cup \mathcal{K}_j|$, and keep an undirected edge *iff* the relation is mutual in the top-k neighbourhood, weighting edges by an RBF kernel of d_J. The resulting graph $\mathcal{G}$ has nodes indexed by run and coarse position, with edges encoding cross-run local similarity on a single problem. The neighbourhood parameter k controls only how many local neighbours are eligible before the mutuality filter. Exact cache constants, kernel scale, and size cap are in Appendix B.2; sensitivity sweeps are in Appendices C.1 and C.6.

3.2 BICONNECTED-COMPONENT DECOMPOSITION AND NON-TRIVIAL REGIONS

We treat $\mathcal{G}$ as a graph of *shared slice-level computation*. A *biconnected component* (BCC) is a maximal subgraph that stays connected after removing any single node, and so groups slices that are mutually reachable through redundant paths. An *articulation point* between two BCCs is a structural connector at which the population of slices splits into separately connected regions; whether such a point coincides with a transition in the generated reasoning is assessed from the corresponding slice text (Appendix E.3). We decompose $\mathcal{G}$ component-wise into BCCs and split the result into

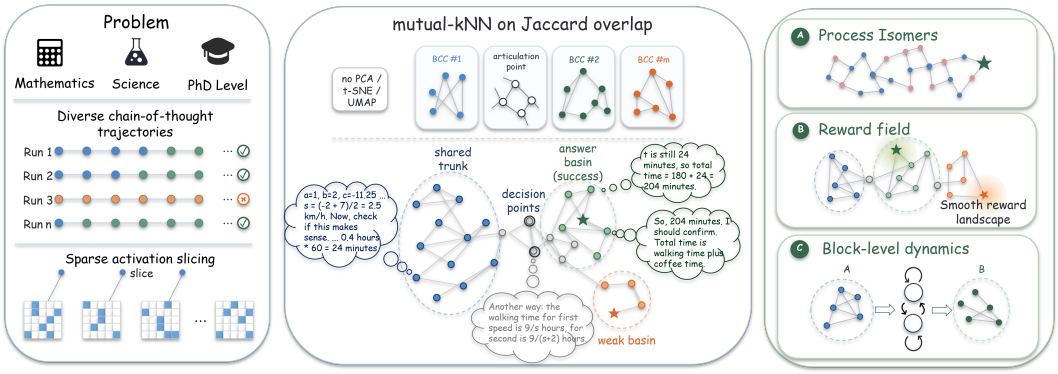


Figure 2: **SliceGraph pipeline.** SliceGraph turns many sampled CoTs for one problem-model cell into an activation-key atlas. Route families are structural units read from the fixed graph without correctness labels; the label-seeded reward field and family-specific navigation kernels are secondary overlays that add answer labels and temporal transitions on top of that scaffold.

trivial K_2 *bridges* and non-trivial $|b|\geq 3$ blocks. No largest-connected-component filter is applied: disconnected components remain eligible, while K_2 bridges are retained as connectors but are not treated as semantic regions. Sensitivity to this convention is reported in Appendix C.1.

We then assign each block one of five roles, namely SHARED_TRUNK, ANSWER_BASIN, WEAK_BASIN, DECISION_POINT, or INTERMEDIATE, based on its run coverage, size, dominant-answer purity, and median position. The explicit thresholds and bridge-versus-cycle formulation are in Appendix B.3; role sensitivity is in Appendix C.1. Roles are interpretive overlays that may use final-answer purity; they do not define route families or isomers. The correspondence between these units and human notions of shared reasoning state is validated in Section 4.2.

3.3 Structural route families and structural process isomers

Two runs that visit similar sets of regions share a route footprint, and may correspond to the same high-level strategy. We turn this intuition into a clustering by reducing each run to the set of non-trivial blocks it visits across all analysed slices, $\mathcal{S}_r = \bigcup_t \mathcal{C}^{\mathrm{nt}}_{r,t}$, and forming a weighted Jaccard with rare-block upweighting $\omega_b = 1/\max(c_b, 0.01)$, where c_b is the fraction of runs reaching block b:

$$J(r, r') = \frac{\sum_{b \in \mathcal{S}_r \cap \mathcal{S}_{r'}} \omega_b}{\sum_{b \in \mathcal{S}_r \cup \mathcal{S}_{r'}} \omega_b}. \tag{1}$$

Repeated visits and visitation order are discarded: families compare which regions a run visits, not how long or in what order it dwells in them (order-sensitive readouts belong to the typed-state overlay). After block-coverage weights are fixed from the full cell, the structural-isomer analysis constructs this run-level co-visitation graph on normalized-correct runs only, keeps edges with $J \geq 0.05$, and partitions it by Louvain community detection at resolution 1.0 and seed 42; we call this the *reference configuration* and justify it independently of the resulting rate in Section 4.4. Incorrect runs define full-cell block coverage, the reward overlay, and visual overlays, but they are not nodes in F_{correct}. Whether two runs placed in the same family are also judged by humans to use the same strategy is the second annotation study of Section 4.2. These families anchor the *structural process isomer* relation:

Definition 1 (Structural process isomer)**.** *Fix a sampled cohort, its SliceGraph atlas, and a family configuration θ (edge threshold, block weighting, Louvain resolution, and seed). Two correct CoT runs r, r' for the same problem-model cell are* structural process isomers *if they reach the same normalized final answer and are assigned to different families in $F_{\mathrm{correct}}(\theta)$.*

The relation is defined relative to the sample, atlas, and configuration, as is standard for clusters fitted in a representation space; a family boundary does not by itself establish a different high-level solution strategy. Section 4.4 reports how the relation moves under θ, under resampling, and under held-out atlases.

3.4 SECONDARY OVERLAYS: REWARD FIELD AND TYPED-STATE DYNAMICS

Two descriptive layers are computed on the fixed scaffold after the families are set. The *label-seeded reward field* seeds each block with its support-shrunk excess success rate and diffuses the seed over block-cut and temporal adjacency for 24 PageRank-style steps (α=0.65); the top quartile of its strictly positive values is the *high-value core*, and a cell whose core has several connected components is *multi-core* (Appendix D, Eqs. 6–11). Because the seed uses correctness labels, the field is an outcome-conditioned overlay: it is never used to build families or isomers, and its component counts are geometric descriptors rather than evidence of privileged solution regions (Section 4.6). The *typed-state overlay* lifts each run to a chain of states $(\text{role}, \text{posbin}, \text{core})$ over its primary blocks, estimates a Laplace-smoothed transition kernel P_f per family, and compares families by row-averaged total variation against a size-preserving family-label shuffle (Appendix D.2, Eqs. 12–13).

SliceGraph is evaluated as a measurement object for process geometry; Section 4.5 tests what it supports as a retrospective audit, and no online use inside decoding is claimed.

4 EXPERIMENTS

4.1 SETUP

We evaluate SliceGraph on five math and science reasoning benchmarks: AIME24 (Zhang & Math-AI Team, 2024), AIME25 (Zhang & Math-AI Team, 2025), BRUMO25 and HMMT25 from MathArena (Balunović et al., 2025), and GPQA Diamond (Rein et al., 2023). The primary corpus uses three 4B and 8B models sampled with chain-of-thought prompting: DeepSeek-R1-0528-Qwen3-8B (**R1-8B**), Qwen3-4B-Thinking-2507 (**Think**), and Qwen3-4B-Instruct-2507 (**Inst**), all run at the canonical configuration. The primary corpus contains 60,622 graph-valid analysed trajectories from 61,056 sampled generations (4 incomplete cache records and 430 graph-invalid runs are dropped; Appendix B.1) across 954 problem-model cells, i.e. 318 unique benchmark problems each evaluated under the three primary models at N=64 runs per cell. N=64 is an observation budget rather than an asymptotic regime: it retains 95.0% of the route-family coverage observed on deeper N=256 caches, and block coverage saturates near $N\approx$32 while family and core counts keep growing (Appendices C.1 and C.7). We run a cross-architecture replication on DeepSeek-R1-Distill-Llama-8B (**Llama-8B**), a within-family cross-scale replication on Qwen3-32B (**Qwen3-32B**), and a cross-generation cross-scale replication on the math-specialized Qwen2.5-Math-72B-Instruct (**Q2.5-72B**), with per-model results in Appendices F.3, F.2, and F.1. Effect sizes and uncertainty conventions are detailed in Appendix B.4. The reward, multi-core, typed-state, and family-TV analyses use progressively stricter subsets carved from the canonical 954 problem-model cells, with the filtering chain in Appendix B.1. Correctness and same-answer membership are evaluated after benchmark-specific answer normalization, and process-isomer comparisons are made only within the normalized correct-answer class. Graph analysis is CPU-only (2–5 s per cell at N=64, under 4 GB; Appendix B.4).

We organise the experiments as a validity ladder: semantic validation of the graph units (§4.2), representation sensitivity (§4.3), calibrated route multiplicity (§4.4), a consensus-reliability audit (§4.5), and the secondary overlays with cross-model replication (§4.6). Each question is paired with a targeted control: human annotation, representation substitution, configuration sweeps with non-singleton, split-half, and held-out diagnostics, label-free family construction with correctness used only to evaluate the audit, and label-shuffle and permutation nulls for the overlays (Table 17 in the appendix).

4.2 **RQ1.** ARE THE GRAPH UNITS SEMANTIC?

Before reading geometry off the graph, we ask whether its two units, biconnected blocks and structural route families, line up with human notions of shared reasoning state and shared strategy. Two blinded annotation studies, each run by two independent non-author annotators on matched within-unit and across-unit pairs, give a two-part answer: strong within-unit coherence, and only partial correspondence between cross-family boundaries and human-named strategies.

Biconnected blocks correspond to shared reasoning state. We constructed 72 blinded slice-text pairs, half within the same BCC and half length-matched across BCCs, and asked two independent

annotators to judge whether the two slices represent the same reasoning operation under matched strict execution-level rubrics. Annotator A matches the within/across-BCC labels on 88.9% of pairs and Annotator B on 90.3%; the two annotators agree with each other on 69 of the 71 pairs on which both are decisive ($\mathbf{97.2}\%$). The annotation protocol and agreement statistics are reported in Appendices E.2 and E.3.

Families are within-family coherent; cross-family boundaries are finer than strategies. A separate 77-pair pack drawn from 8 deliberately selected family-rich cases, with 46 within-family and 31 length-matched across-family pairs, asked whether two runs use the same high-level strategy. Annotator A judges 34 of 36 decisive within-family pairs ($\mathbf{94.4}\%$; $34/46{=}73.9\%$ of all within-family pairs) to use the same strategy; cross-annotator agreement is $\kappa{=}0.74$ on the across-family slice and 81.8% exact agreement on the 55 pairs where both annotators are decisive (Appendix E.4). Among the 31 across-family pairs, 10 are judged different-strategy, 11 same-strategy, and 10 uncertain: a 47.6% different-strategy rate among decisive pairs, or 32.3% when uncertainty is counted as not establishing a difference. These are conditional validation rates on family-rich cases, not a corpus-wide prevalence of strategy-level isomerism, and we do not multiply them by the structural rate below. A cross-family boundary may therefore mark a human-named strategy change, a finer difference in intermediate operations, or a representation or partitioning effect. Hence *structural* process isomerism denotes graph-measured route differences, and we reserve "different strategy" for human-judged pairs.

4.3 **RQ2.** How does the slice representation affect the scaffold?

If the scaffold distorts process geometry, the isomer claims that follow are built on artefactual structure. We hold the mutual-kNN framework, BCC decomposition, and family construction fixed and vary only the slice representation, evaluating each by two criteria: whether it yields a valid non-trivial BCC scaffold, and how much correct-family multiplicity the resulting partition retains.

Evidence. Under the fixed scaffold, activation-key Jaccard yields the highest correct-family multiplicity across primary models (Table 1; per-model breakdown in Appendix C.2). Cosine and overlap on the same sparse keys are close, supporting the activation-key object rather than a single metric artefact. PCA-50 remains graph-valid but compresses family counts by $\sim$1 per cell, while slice-level MiniLM nearly collapses the atlas: only $21/954$ cells yield a valid decomposition, with at most 11 valid cells for any primary model.

Table 1: Representation ablation inside the fixed SliceGraph scaffold (mutual-kNN, BCC decomposition, and family construction unchanged). Multi-fam rate / mean correct families per valid cell. Canonical rows evaluate all 954 cells; collapsed representations report their valid-cell counts (see Appendix C.2). This is not an end-to-end benchmark of prior pipelines.

Scaffold representation	R1-8B	Think	Inst
Activation keys, Jaccard (ours)	$\mathbf{89.3/5.14}$	$\mathbf{86.2/4.61}$	$\mathbf{81.1/4.21}$
Activation keys, cosine	$88.1/4.80$	$85.5/4.36$	$79.6/3.87$
PCA-50 + cosine	$81.4/3.38$	$81.4/3.33$	$73.8/3.00$
MiniLM slice-text	*collapsed* ($21/954$ valid cells; $\leq$11 per model)		

This ablation isolates representation sensitivity within SliceGraph. Higher family counts alone do not establish better recovery of reasoning structure; semantic agreement (RQ1) and partition stability (RQ3) are evaluated separately. A plausible but unverified explanation for the MiniLM collapse is insufficient within-problem contrast between slices (Appendix C.2). We use activation-key Jaccard as the reference representation.

4.4 **RQ3.** How much route diversity hides behind correct-answer agreement?

Structural process isomerism is a family-defined relation (Definition 1). We first measure how prevalent it is under the reference configuration, then report its configuration envelope and its behaviour under resampling and held-out atlases; the overlays are layered on in §4.6. Figure 3

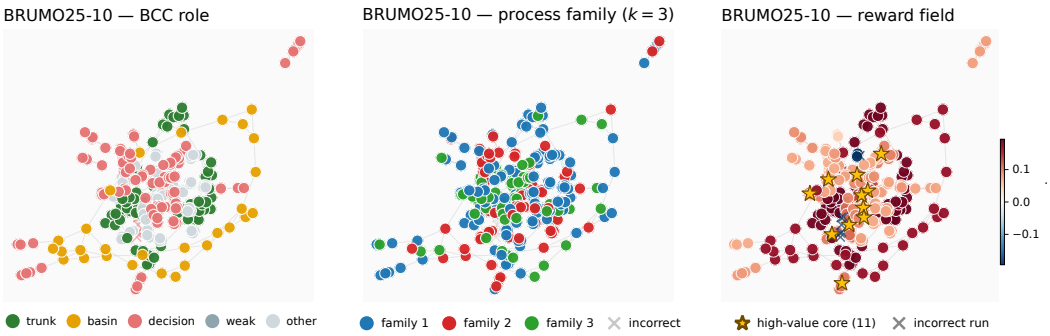


Figure 3: Three readouts of a single SliceGraph for BRUMO25-10 (Inst, 64 runs, 60 correct). Every node is one CoT **slice**; 273 slices in non-trivial biconnected components are shown, with trivial K_2 bridge components omitted. **Left:** BCC block-type role (trunk, basin, decision, weak, other). **Middle:** correct-only route families (3 families; $\times$ = incorrect runs). **Right:** propagated reward field seeded by per-block correctness deviation; ★ marks each high-value-core component (11 cores). Family colours and core locations split differently: route families and the reward overlay are related but non-identical layers.

Table 2: Route multiplicity and its configuration envelope (primary corpus; problem-clustered bootstrap 95% CIs). Fixed-scale configurations vary threshold, weighting, and seed at the reference scale; the 32-configuration envelope also varies scale and mutuality (Appendix C.6).

Readout	Value	Denominator
Cells with $\geq$2 correct-only families	85.5% (816)	954 canonical cells
same, among cells with $\geq$2 correct runs	97.8%	834 eligible cells
Cells with $\geq$2 families of size $\geq$2	92.3% [90.0, 94.4]	eligible cells
Correct runs in singleton families	1.30% [1.11, 1.52]	correct runs
Pairwise isomer rate (reference configuration)	76.6% [75.6, 77.5]	834 cells (per-cell mean)
Pairs separated in $\geq$80% of 9 fixed-scale configurations	52.6% [51.5, 53.7]	correct-run pairs
Pairs separated in $\geq$80% of 32 configurations	14.7%	correct-run pairs

visualises three layers of one atlas; the population ECDF and per-dataset breakdowns are in the appendix (Figure 5a, Table 12).

Route multiplicity. Table 2 collects the population readouts with their denominators. **85.5**% of the 954 canonical cells (816) contain more than one correct-only route family, i.e. 97.8% of the 834 cells with at least two normalized-correct runs. Multiplicity is not carried by isolated runs: **92.3**% of eligible cells contain at least two families with at least two members each, and singleton families cover only 1.30% of correct runs. At the pair level, **76.6**% of normalized-correct run pairs are cross-family on average, with a problem-clustered bootstrap 95% CI of [75.6, 77.5]%.

Configuration dependence. The pairwise rate is a configuration-dependent quantity, so we report it as a reference-configuration estimate with its envelope rather than as an invariant prevalence. Varying one axis at a time gives 74.4–79.3% for the edge threshold (τ=0.025–0.1), 62.0–76.6% for rare-block weighting, 75.2–76.6% for slice granularity, 47.6–82.7% for k=4–12, and 36.5–91.5% for Louvain resolution (ρ=0.5–2.0); the full resolution$\times$threshold grid spans 24.7–91.7% (Appendix C.6). Removing mutuality reduces the rate to 17.5% and the multi-family rate to 35.9%, so reciprocal neighbourhoods are a defining component of SliceGraph rather than a tuning option. Seeds barely matter (ARI$\approx$0.95), whereas changing slice granularity or k preserves aggregate multiplicity but reorganizes which pairs are separated (ARI 0.03–0.16). Holding the scale fixed, 52.6% of correct-run pairs are separated in at least 80% of nine threshold, weighting, and seed settings, with a bimodal distribution (48.4% of pairs separated in 90–100% of settings, 9.7% in at most 20%); across 32 configurations that also change scale and mutuality, the consensus rate is 14.7%: recurring

Table 3: Accuracy (%) of label-free family-aware aggregation versus self-consistency (SC) on the primary corpus, with families from F_{all} and cohorts of N runs subsampled from the canonical caches. Family-aware rules trail SC by 0.19–0.85 points at every N.

N	SC	One family, one vote	$\sqrt{\text{size}}$ weight	Family support
8	**66.37**	66.18	65.88	66.05
16	**66.95**	66.62	66.60	66.44
32	**67.55**	67.31	67.10	66.83
64	**68.28**	67.43	67.99	67.43

multi-route structure is robust at a stated scale, while exact pair identities are scale-relative. The reference configuration is fixed independently of the rate: ρ=1 is the standard Louvain resolution, τ=0.05 the central value of the prespecified sweep, inverse-frequency weighting discounts ubiquitous blocks, and the bridge convention excludes edges that overfragment regions (K_2 bridges move the multi-family rate only to 86.8% but inflate mean families from 4.66 to 11.27; Appendix C.1).

Sampling robustness. Split-half resampling preserves the signal on a stricter split-valid subset: both halves are multi-family in 97.2% of cells (n=779). A held-out atlas built from 32 training runs covers 95.5% of held-out blocks and yields multiple families in 91.0% of eligible held-out cells; exact identities transfer less strongly (held-out pairwise rate 60.1% versus 77.8% for the matched full partition; median NMI 0.24), so the held-out check supports out-of-sample multiplicity and coverage, not one-to-one family recovery (Appendix C.4). Fixing the correct-run count to m=4 still yields 82.1% multi-family cells and a 58.2% isomer rate (Appendix C.5). Low-isomer cells remain, and SliceGraph recovers multi-route structure when present rather than forcing it. Thus the same correct answer is usually not a single route endpoint but a set of recurring, graph-separated route footprints whose exact partition is scale-relative.

4.5 **RQ4.** What does route structure reveal about consensus reliability?

RQ3 measures diversity among correct runs. We now ask whether route structure supports a use that answer counts cannot: auditing how narrowly an apparent consensus is supported.

Protocol. Families are fitted to all sampled runs of a cell without correctness labels (F_{all}, Section 3). For a cohort of N runs we take the self-consistency majority answer and call it *route-collapsed* when all of its supporting runs belong to one structural family. Correctness is used only to evaluate the flag. Cohorts with N<64 are subsampled from the canonical caches, and their block regions come from the amortized unlabeled 64-run atlas, so the analysis is retrospective and does not establish online compute savings (Appendix D.7).

Family reweighting does not beat answer frequency. Three label-free family-aware aggregation rules (one vote per family, $\sqrt{\text{size}}$-weighted family votes, and family-support scoring) trail self-consistency by 0.19–0.85 percentage points at every N (Table 3). Route families should therefore complement, not replace, answer frequency.

Route-collapsed consensus is less reliable. At N=8 the flag marks **7.8**% of cohorts; their accuracy is **33.6**%, versus 69.2% for the remaining cohorts and 66.4% overall. At N=16 and 32, within vote-share strata that contain both groups, majorities supported by several families are 14.6–44.4 points more accurate than single-family majorities. Matched within-problem comparisons are mostly null, and family support does not improve global AUROC: the demonstrated signal is a problem-level flag for structurally narrow consensus, not a way to select a different answer within the same problem.

What the measurement is for. Two uses are demonstrated: measuring whether agreement repeats one route or spans several, with localized trunks, forks, and rejoin regions (Figure 3, Appendix D.6), and flagging route-collapsed consensus for review. Selecting one representative trace per family for manual audit, testing whether extra samples discover or repeat routes (Appendix C.7), and curating route-diverse training data use the same output but are not evaluated here.

4.6 SECONDARY OVERLAYS AND CROSS-MODEL REPLICATION

Label-seeded reward overlay. On the 479 reward-evaluable cells, 67.6% are multi-core, the family–core footprint specialization is 0.59 on the 324 multi-core cells, and family and core counts correlate weakly (Spearman ρ=0.08; Figure 5b,c in the appendix). These are descriptors of the constructed field, not evidence of privileged solution regions: under label permutation (100 shuffles) the multi-core rate and specialization are *higher* (85.4% and 0.74) because scattered positive seeds fragment, and field sharpness equals its permutation range (1.55×, z=0.08). The one overlay result that exceeds its matched null is that family-conditioned cores overlap the global core more than under permutation (Jaccard 0.17 vs. 0.10 [0.097, 0.105], z=29.4), an outcome association conditional on this construction (Appendix D.4). A disjoint-core criterion adds only 1.3 pp to the family-isomer relation.

Family-specific dynamics. On the 403-cell core-eligible multi-family subset, row-averaged family-TV exceeds the family-label-shuffle p_{95} in 80.9% of cells (median z=3.14); restricting to states visited by both families (common-support TV) preserves the direction in 77.4% of cells (median z=3.45, median 7 shared states), and family-TV is only weakly correlated with footprint specialization (−0.07). Families thus also differ in navigation under this null, with the caveat that they are themselves defined by footprints (Appendix D.2, Figure 6).

Cross-model replication. Llama-8B (architecture transfer) and Qwen3-32B (within-family scale) reproduce the route-family pattern (78.3% and 89.6% multi-family cells) and the family-TV result; Qwen2.5-Math-72B attenuates multiplicity on the math benchmarks (67.3% overall) while matching the pattern on GPQA (86.4%). Because these models differ in training, architecture, and task performance, the comparison does not isolate a scale effect or establish its cause (Appendices F.3, F.2, F.1).

5 CONCLUSION

SliceGraph measures the structural route support behind answer agreement in multi-run reasoning: it maps sparse activation keys from many CoTs into a graph of shared and diverging route footprints and reads structural route families from it. On the studied models and benchmarks, correct traces that agree on an answer usually occupy several recurring activation-defined routes. Human annotation supports within-family coherence, while configuration and held-out analyses show that exact family boundaries are scale- and sample-dependent: the reported rates are calibrated measurements, not a taxonomy of strategies. As a use, structurally narrow consensus marks a small, less reliable subset of cohorts, whereas family-based vote reweighting does not improve answer selection. Its value for prospective abstention, sampling, or trajectory selection remains to be established under protocols that use only decision-time information.

6 LIMITATIONS AND BROADER IMPACT

SliceGraph inherits the limits of observational activation analysis: it requires hooks and cached sparse activations (cost in Appendix B.4). Family identity depends on the sampled cohort, slice granularity, graph construction, and clustering resolution; the reported separation rate is a reference-configuration measurement with a stated envelope, and N=64 is an observation budget under which family and core counts are still growing. The human family audit uses eight deliberately selected family-rich cases and supports conditional semantic alignment, not corpus-wide strategy diversity. The route-collapse audit uses block regions from a full 64-run atlas, so smaller cohorts establish retrospective analysis, not reduced sampling cost or early stopping. The reward overlay is descriptive and label-seeded, so label permutations are stress tests of that layer rather than one-sided nulls (Appendix D.4). Claims are scoped to sampled textual CoT on non-code math and science tasks; tools, APIs, and agent actions need modality-specific representations and validation. Broader impacts are diagnostic, but selection or early-stopping use could amplify sampling biases or suppress rare correct routes.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods (AI coding assistants helped write the activation-cache, graph-construction, and analysis code), for cleaning and reformatting data, for interpreting results, and for the qualitative analysis of annotation cases. We have not used generative AI tools for producing synthetic data, building theoretical models or conceptual frameworks, formulating mathematical claims or proofs, proposing or refining hypotheses, designing the methodology or experiments, or translation. Additionally, we used generative AI tools for tasks with recommended disclosure: searching and summarizing related literature, editing and polishing the text, and converting the LaTeX sources of this paper to the ICLR 2027 template. We have reviewed all AI-assisted work: AI-assisted code was verified by tests and by reproducing the reported numbers, AI-assisted data processing and result interpretation were re-checked by the authors against the experiment outputs, and all AI-edited text was checked sentence by sentence by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes open-weight language models (DeepSeek-R1-0528-Qwen3-8B, Qwen3-4B-Thinking/Instruct-2507, DeepSeek-R1-Distill-Llama-8B, Qwen3-32B, and Qwen2.5-Math-72B-Instruct) on public benchmarks (AIME 2024/2025, BRUMO 2025, HMMT 2025, and GPQA Diamond), and releases no new dataset. The two blinded annotation studies (Appendix E.2) were carried out by two non-author members of our research group who volunteered without compensation and gave informed consent; they judged pairs of model-generated reasoning text only, no personal data was collected, and the study therefore did not require institutional review. SliceGraph is a post-hoc measurement object rather than a decoder; the risks of using it for selection or early stopping are discussed in Section 6.

REPRODUCIBILITY STATEMENT

The graph construction, block decomposition, route families, reward field, and typed-state dynamics are specified in Section 3, and the models, benchmarks, and corpus sizes in Section 4.1. The appendix gives the analysis subsets and their filtering chain (Appendix B.1), the activation-cache pipeline and all graph hyperparameters (Appendix B.2), the block-role thresholds (Appendix B.3), statistical conventions (Appendix B.4), the reward-field construction (Appendix D), the consensus-audit protocol (Appendix D.7), the null constructions (Appendices D.4 and D.5), hyperparameter sensitivity (Appendices C.1 and C.6), and the annotation protocol (Appendix E.2). All models and datasets are publicly available. Anonymized code and the exported headline tables are submitted as supplementary material, and the code will be released publicly upon publication.

REFERENCES

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025. doi: 10.48550/arXiv.2505.23281. URL https://arxiv.org/abs/2505.23281.

Maciej Besta, Nils Blach, Aleš Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024. doi: 10.1609/aaai.v38i16.29720. URL https://ojs.aaai.org/index.php/AAAI/article/view/29720.

Kang Chen, Yaoning Wang, Kai Xiong, Zhuoka Feng, Wenhe Sun, Haotian Chen, and Yixin Cao. Do LLMs signal when they're right? evidence from neuron agreement. *arXiv preprint arXiv:2510.26277*, 2025. doi: 10.48550/arXiv.2510.26277. URL https://arxiv.org/abs/2510.26277.

Kang Chen, Zhuoka Feng, Sihan Zhao, Kai Xiong, Junjie Nian, Yaoning Wang, Changyi Xiao, and Yixin Cao. NEX: Neuron explore–exploit scoring for label-free chain-of-thought selection and

model ranking. *arXiv preprint arXiv:2602.05805*, 2026. doi: 10.48550/arXiv.2602.05805. URL https://arxiv.org/abs/2602.05805.

Dongkyu Cho, Amy B. Z. Zhang, Bilel Fehri, Sheng Wang, Rumi Chunara, Hengrui Cai, and Rui Song. Correct reasoning paths visit shared decision pivots. *arXiv preprint arXiv:2509.21549*, 2025. doi: 10.48550/arXiv.2509.21549. URL https://arxiv.org/abs/2509.21549.

Hamed Damirchi, Ignacio Meza De la Jara, Ehsan Abbasnejad, Afshar Shamsi, Zhen Zhang, and Javen Shi. Truth as a trajectory: What internal representations reveal about large language model reasoning. *arXiv preprint arXiv:2603.01326*, 2026. doi: 10.48550/arXiv.2603.01326. URL https://arxiv.org/abs/2603.01326.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. LLM reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*, 2026. doi: 10.48550/arXiv.2604.05655. URL https://arxiv.org/abs/2604.05655.

Xue Wen Tan, Nathaniel Tan, Galen Lee, and Stanley Kok. The shape of reasoning: Topological analysis of reasoning traces in large language models. *arXiv preprint arXiv:2510.20665*, 2025. doi: 10.48550/arXiv.2510.20665. URL https://arxiv.org/abs/2510.20665.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zihao Wei, Liang Pang, Jiahao Liu, Wenjie Shi, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Fei Sun, Huawei Shen, and Xueqi Cheng. The evolution of thought: Tracking LLM overthinking via reasoning dynamics analysis. *arXiv preprint arXiv:2508.17627*, 2025. doi: 10.48550/arXiv.2508.17627. URL https://arxiv.org/abs/2508.17627.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Mapping the minds of LLMs: A graph-based analysis of reasoning LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17751–17763, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.896. URL https://aclanthology.org/2025.emnlp-main.896/.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2024. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/math-ai/aime24.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

Tianjun Zhong, Linyang He, and Nima Mesgarani. From chains to DAGs: Probing the graph structure of reasoning in LLMs. *arXiv preprint arXiv:2601.17593*, 2026. doi: 10.48550/arXiv.2601.17593. URL https://arxiv.org/abs/2601.17593.

Zhanke Zhou, Zhaocheng Zhu, Xuan Li, Mikhail Galkin, Xiao Feng, Sanmi Koyejo, Jian Tang, and Bo Han. Landscape of thoughts: Visualizing the reasoning process of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XpoQ812d0A. Poster.

APPENDIX ROADMAP

- **Glossary**: standard versus introduced terms (§A).
- **Implementation and count conventions**: analysis subsets (§B.1); cache pipeline and graph hyperparameters (§B.2); block roles and bridge-vs-cycle convention (§B.3); reproducibility and statistics (§B.4).
- **Representation and family robustness**: block-scale and hyperparameter sensitivity (§C.1); representation ablation per pipeline×model (§C.2); fixed-scaffold split-half stability (§C.3); held-out atlas validation (§C.4); correct-run controlled isomer rate (§C.5); family-construction sensitivity (§C.6); isomer-discovery curves under run subsampling (§C.7).
- **Process-isomer geometry and dynamics**: reward-field construction (§D); multi-core / multi-family population summary (§D); per-model block-type localisation (§D.1); family-specific navigation diagnostics (§D.2); typed-state sensitivity and scope (§D.3); null-control ledger and label-permutation reward stress test (§D.4); family-TV null construction (§D.5); worked route report on GPQA-71 (§D.6); route-collapsed consensus and family-aware voting (§D.7).
- **Human annotation protocol**: family detection robustness (§E.1); annotation protocol (§E.2); Task A (BCC shared-state) (§E.3); Task B (route families) (§E.4); qualitative annotation cases (§E.5).
- **Cross-model replications**: Q2.5-72B cross-generation cross-scale (§F.1); Qwen3-32B within-family cross-scale (§F.2); Llama-8B cross-architecture (§F.3).

A GLOSSARY OF TERMS

Table 4 separates standard terminology from the operational terms introduced in this paper and from the secondary descriptive terms used only for the overlays.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 ANALYSIS SUBSETS AND COUNT CONVENTIONS

Unique runs are sampled trajectories before graph construction; *analysed rows* are runs surviving slice selection. The second count is slightly smaller where individual runs yield no valid slice path.

Count ledger. The canonical primary atlas corpus has 318 unique benchmark problems evaluated under three primary models, giving $318 \times 3 = 954$ problem-model cells. At $N{=}64$ generations per cell this yields 61,056 sampled generations; after dropping 4 incomplete cache records, the corpus contains 61,052 runs. Graph-valid filtering further drops 430 runs whose stratified slice sampling yields an empty path, leaving 60,622 analysed rows. No largest-connected-component filter is applied: BCCs are computed component-wise on the selected graph, so disconnected rare routes are not silently removed. We reuse "954-cell corpus" as shorthand for these 954 problem-model cells. The reward-evaluable subset is stricter: after graph-valid filtering a cell must have at least 3 correct and 3 non-correct runs, at least two non-trivial blocks with non-empty run–block sequences, at least one seed-eligible block ($3 \le n_b \le n - 3$ visitors), and non-empty strictly positive support in $\mathbf{v}^{(24)}$. In the exported 479 reward-evaluable cells the minimum observed counts are 3 correct runs, 3 non-correct runs, 4 non-trivial blocks, 3 seed-eligible blocks, and one positive high-value-core block.

Derivation chain. The main atlas subsets form two refinement chains. The route-family chain is

$$954 \supset 834 \supset 816,$$

Table 4: Standard terms, terms introduced by this paper, and secondary descriptive terms.

Term	Meaning
Standard terms	
Chain-of-thought (CoT)	A generated trace containing intermediate reasoning before the final answer.
Self-consistency	Sample several CoTs and return the most frequent normalized final answer.
Jaccard similarity; mutual-kNN; biconnected component (BCC); articulation point; Louvain	Standard set-similarity, graph-construction, graph-decomposition, and community-detection notions.
Post-hoc	Analysis performed after generation on completed trajectories.
Within-run / whole-run	Analysis inside one trace / of one complete trace, as opposed to across runs.
Terms introduced here	
Problem–model cell	One problem evaluated with one model and its N sampled CoTs.
Slice	A 256-token segment of one CoT, represented by its sparse activation-key set.
SliceGraph (atlas)	The post-hoc mutual-kNN graph over the slices of one cell.
Route block	A non-trivial ($\lvert b\rvert \geq 3$) BCC of the SliceGraph; K_2 bridges are connectors, not blocks.
Structural route family	A group of runs with similar block-visitation footprints (Louvain on weighted Jaccard); *within-family* = same group.
F_{correct} / F_{all}	Family partition fitted on normalized-correct runs only / on all sampled runs without labels.
Structural process isomer	Two correct runs with the same normalized answer in different families of $F_{\text{correct}}(\theta)$.
Reference configuration	`sep_up`=8, k=6, mutual-kNN, τ=0.05, inverse-frequency weighting, ρ=1, seed 42.
Route-collapsed consensus	A self-consistency majority whose supporting runs all belong to one family of F_{all}.
Secondary descriptive terms	
Label-seeded reward field	Outcome-conditioned overlay obtained by diffusing per-block excess success rates over the fixed scaffold.
High-value core; core footprint	Top quartile of the positive field; the set of core components a run visits.
Typed-state transition analysis; family-TV	Transitions among (role, phase, core) states of the primary blocks; row-averaged total variation between family kernels.

namely canonical corpus $\supset$ cells with at least two normalized-correct runs (the denominator of the pairwise isomer rate) $\supset$ cells with at least two correct-only families; the split-half-valid subset (779) additionally requires two correct runs in each half. The overlay chain is

$$954 \;\rightarrow\; 479 \;\rightarrow\; 403,$$

namely canonical corpus $\rightarrow$ reward-evaluable cells $\rightarrow$ core-eligible multi-family cells, with the 324 multi-core cells a further subset of the 479. The alignment readouts in Table 13 then use natural denominators: for example, specialization is averaged on the 324 multi-core reward-evaluable cells, while family-TV uses the 403-cell subset.

B.2 Activation-Cache and Graph Implementation

Sparse activation logging. We record post-gating MLP intermediates under the standard SwiGLU/GeGLU convention used by Qwen3, DeepSeek-R1-Distill-Qwen, DeepSeek-R1-Distill-Llama, and Qwen2.5-Math. Let $u_{\ell,t}, g_{\ell,t} \in \mathbb{R}^{d_{\text{ff}}}$ denote the FFN "up" and "gate" pre-activations at layer ℓ and token t, and let $\phi(\cdot)$ denote the gating nonlinearity (SiLU here). We score each neuron by the positive part of its post-gating contribution,

$$h_{\ell,t} \triangleq \phi(g_{\ell,t}) \odot u_{\ell,t}, \qquad a_{k,t} \triangleq \max\big(0, (h_{\ell,t})_j\big), \tag{2}$$

with each neuron keyed across layers as

$$k = (\ell \ll 16) \;\big|\; j. \tag{3}$$

At each token we retain the top `global_topk`= 2000 neurons by $a_{k,t}$.

Table 5: Main analysis subsets used throughout the atlas version of the paper.

Analysis	Cells	Runs / rows	Notes
Canonical primary corpus	954	60,622 analysed rows	3 primary models, 5 non-code datasets, canonical `sep_up`=8, k=6, mutual-kNN
Pairwise isomer-rate subset	834 problem-model cells	correct-run pairs	Cells with $\geq$2 normalized-correct runs after graph-valid filtering; denominator of the 76.6% rate (816/834=97.8% are multi-family)
Multi-family subset	816 problem-model cells	cell-level subset	Cells with $\geq$2 correct-only route families; denominator of the coverage-loss readout
Split-half-valid subset	779 problem-model cells	cell-level subset	Cells with $\geq$2 correct runs in both deterministic halves (§C.3)
Reward-evaluable overlay subset	479 problem-model cells	cell-level subset	Requires $\geq$ 3 correct and $\geq$ 3 non-correct runs, $\geq$ 2 non-trivial blocks, at least one seed-eligible block, and non-empty positive field support
Core-eligible multi-family subset	403 problem-model cells	family-level kernels	Multi-family cells with at least one high-value core; used for family-TV and joint family/core checks; alignment readouts use natural denominators in Table 13
Multi-core subset	324 problem-model cells	cell-level subset	Reward-evaluable cells with $\geq$2 high-value-core components; denominator of specialization
Consensus-audit cohorts	canonical cells	$N \in \{8, 16, 32, 64\}$ subsampled cohorts	Families from F_{all}; block regions from the 64-run atlas (§D.7)
Cross-architecture replication	258	16,512 analysed rows	DeepSeek-R1-Distill-Llama-8B on AIME24, AIME25, and GPQA
Within-family cross-scale replication	318	20,352 analysed rows	Qwen3-32B (thinking mode) on the full 5-dataset non-code suite
Cross-generation cross-scale replication	318	20,196 analysed rows	Qwen2.5-Math-72B-Instruct on the full 5-dataset non-code suite
Process-family semantics (annotation)	8 cases	102 runs	46 within-family + 31 across-family judged pairs; two independent non-author annotators
BCC reasoning-state semantics (annotation)	36 cells	72 slice pairs	36 within-BCC + 36 matched across-BCC pairs; two independent non-author annotators

Window aggregation and slice keys. Tokens are grouped into rows of `slice_size`= 32. For row r with token-index set T_r, we form $A_{k,r} = \sum_{t \in T_r} a_{k,t}$ and keep the `slice_topk`= 500 neurons with largest mass:

$$\mathcal{K}_r = \text{top}_{K=500}\{k : A_{k,r}\}, \qquad |\mathcal{K}_r| \leq K. \tag{4}$$

The analysis pipeline then aggregates every consecutive `sep_up`=8 rows into one analysed slice, so one analysed slice summarises 256 tokens of CoT.

Graph construction. Canonical SliceGraph construction uses mutual-kNN with k=6, distance scale σ=0.35, and size cap M=2,600 slices per problem. Problems exceeding M are stratified-subsampled over runs and positions before graph construction. Edge weights are RBF transforms of Jaccard distance on the sparse key sets.

B.3 Block Roles, Thresholds, and Bridge-vs-Cycle Convention

The five block roles used in §3.2 are assigned from each block's run coverage c_b, size $|b|$, dominant-answer purity π_b, unique run count $n_{\text{runs}}(b)$, and median position $\text{medpos}(b)$:

$$\text{role}(b) = \begin{cases} \text{SHARED_TRUNK} & c_b > 0.4,\ |b| \geq 6,\ \text{medpos}(b) \leq q_{0.5}^{\text{trunk}} \\ \text{ANSWER_BASIN} & |b| \geq 6,\ \text{medpos}(b) > q_{0.5}^{\text{basin}},\ \pi_b \geq 0.6,\ n_{\text{runs}}(b) \geq 3 \\ \text{WEAK_BASIN} & |b| \geq 6,\ \text{medpos}(b) > q_{0.5}^{\text{basin}} \text{ but answer-basin criteria fail} \\ \text{DECISION_POINT} & |b| \leq 5 \text{ and touches an articulation point} \\ \text{INTERMEDIATE} & \text{otherwise.} \end{cases} \quad (5)$$

Bridge-vs-cycle convention. The biconnected decomposition partitions the graph into trivial K_2 bridge blocks and non-trivial blocks of size $|b| \geq 3$ that are 2-vertex-connected. All main-text constructs operate on the non-trivial subset because including bridges inflates length-driven fragmentation while adding little semantic content to the atlas.

Primary-block tie breaks. When a slice is covered by multiple non-trivial blocks, the primary block $b^{\star}_{r,t}$ is chosen by largest block size in nodes, with ties broken by larger unique-run count, earlier minimum slice position, then smaller block id. Slices with no non-trivial block are skipped in typed-state trajectories rather than mapped to a sentinel.

B.4 Reproducibility and Statistics

Fresh CoT generation uses framework-level nondeterminism rather than fixed per-run sampling seeds, so bitwise reruns of the analysis are supported by the generated cache and graph summaries; the paper and the code in the supplementary material provide full algorithmic details and exported headline tables. The paper reports problem-clustered bootstrap intervals (B=1000) for headline proportions and means where appropriate, and null-controlled z-scores for graph or transition-structure diagnostics. Multiple related null-based tests within the same experiment family use Benjamini–Hochberg FDR control.

Model checkpoints. All measured systems are open-weight Hugging Face checkpoints (Table 6); no fine-tuning is applied and all generations are zero-shot CoT with each model's default chat template.

Table 6: Model checkpoints used for CoT generation.

Paper name	HuggingFace ID	Mode / setting	Arch.
R1-8B	`deepseek-ai/DeepSeek-R1-0528-Qwen3-8B`	thinking-only	Qwen3
Think	`Qwen/Qwen3-4B-Thinking-2507`	thinking-only	Qwen3
Inst	`Qwen/Qwen3-4B-Instruct-2507`	non-thinking-only	Qwen3
Llama-8B	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	thinking-only	Llama-3.1
Qwen3-32B	`Qwen/Qwen3-32B`	`enable_thinking=True`	Qwen3
Q2.5-72B	`Qwen/Qwen2.5-Math-72B-Instruct`	default	Qwen2.5

Hardware and compute. CoT generation: 8× NVIDIA H20-96GB GPUs, using flash-attention-2; the 72B model was served with 8-way tensor parallelism, smaller models on a single H20. Estimated total generation cost: ∼200 GPU-hours across all models and benchmarks. Activation-cache extraction ran on the same hardware during generation. The SliceGraph analysis pipeline (graph construction, BCC decomposition, reward field, and typed-state diagnostics) runs entirely on CPU and completes in 2–5 s per problem at N=64 runs with under 4 GB of peak memory. The sparse activation caches for the full study occupy approximately 150 GB; upstream sampling and activation logging, not the graph analysis, dominate the cost.

Software versions. Python 3.12; PyTorch 2.6; Transformers 5.3; NetworkX 3.6 (BCC decomposition via `biconnected_components`, Louvain via `louvain_communities`); NumPy 2.2; SciPy 1.17; scikit-learn 1.8 (PCA ablation); `sentence-transformers` (MiniLM ablation).

Table 7: CoT generation protocol. We intentionally use a uniform sampling protocol across all measured models for atlas comparability, rather than each model card's recommended decoding parameters; model-specific chat templates and mode switches are still respected.

N per problem-model cell	64
Temperature	0.7
Top-p	0.9; uniform across all models
Top-k	disabled
`max_tokens`	32,768
Chat template	each model's default chat template applied via `tokenizer.apply_chat_template`; no system prompt
User prompt (math)	`Return your final response within \boxed{}.\n Question:\n{problem}`
User prompt (GPQA)	`Please answer the following multiple choice question:\n\n Question: {question}\n\n Options:\nA. {A} B. {B} C. {C} D. {D}\n\n Return your final response within \boxed{} and only include the letter choice (A, B, C, or D) as your final response.`
Answer extraction	parse last `\boxed{...}` via brace-matched extraction; numeric benchmarks verified with `math_verify` (symbolic equivalence); GPQA verified by exact single-letter match
Sampling randomness	framework-level nondeterminism; analysis reruns use distributed caches
Mean / median output tokens	10,208 / 7,578
Length-truncated runs	$4{,}210/61{,}052 = 6.9\%$ (primary corpus; retained, counted as non-correct if no answer extracted)

Random seeds. Louvain community detection uses seed 42; problem-clustered bootstrap uses B=1,000 resamples with `numpy.random.seed(0)` per bootstrap call; graph-null rewiring uses 3 independent seeds per variant. Temperature-sampled CoT runs have no fixed seed (each run is an independent sample from the model's output distribution).

C Representation and Family Robustness

C.1 Block-Scale Sensitivity and Hyperparameters

The paper uses the graph-theoretic split between trivial BCCs (K_2, bridges) and non-trivial biconnected blocks ($|b| \geq 3$), i.e. 2-vertex-connected cyclic regions. Moving from the full decomposition (m=2) to the non-trivial convention (m=3) changes the object substantially: the mean retained block count drops from 575.2 to 49.5. Even so, the family-level atlas picture is stable. Including bridges shifts the multi-family headline only slightly ($85.5\% \rightarrow 86.8\%$) but overfragments the partition, raising mean correct-only families per cell from 4.66 to 11.27 and the between-family length-variance ratio from 0.460 to 0.697. We therefore retain bridges as connectors but not as semantic regions.

Configuration audit. A 32-configuration audit spanning slice granularity (`sep_up`), k, mutuality, edge threshold, rare-block weighting, Louvain resolution, and seed separates ordinary analysis choices from changes of measurement scale. The Louvain seed has little effect (ARI$\approx$0.95 between partitions); threshold, weighting, and resolution alter the partition to varying degrees (single-axis ranges in Table 11); changing slice granularity or k preserves aggregate multiplicity while reorganizing which pairs are separated (ARI 0.03–0.16). Removing mutuality reduces the pairwise isomer rate to 17.5% and the multi-family-cell rate to 35.9%, so reciprocal neighbourhood structure is a defining component of SliceGraph rather than an interchangeable tuning option. The default configuration is therefore not a narrow optimum for aggregate multiplicity, but exact family identity is scale-relative (§C.6). On run-number sufficiency, N=64 retains 95.0% of the route-family coverage observed on deeper N=256 caches.

RBF kernel scale σ. The edge-weight scale $\sigma \in \{0.20, 0.35, 0.50\}$ has no effect on topology-dependent readouts because the mutual-kNN construction determines graph topology independent of

weight scaling: σ modulates edge strength but not edge existence. On the sweep-eligible subset, the multi-family rate remains 97.8% and the pairwise isomer rate remains 76.6% at all three values.

Reward-field diffusion α and core quartile q. Varying the PageRank damping $\alpha \in \{0.50, 0.65, 0.80\}$ and core extraction percentile $q \in \{0.50, 0.75, 0.90\}$ preserves the correct-run field advantage in all configurations: non-canonical variants produce equal or larger correct-vs-incorrect separation on field-score and entry-progress metrics across all three models. The canonical α=0.65, q=0.75 is therefore a *conservative* choice; no variant degrades the reward-field signal or inverts the multi-core structure.

Role thresholds. An OAT sweep over 8 role parameters (46 configurations including default) shows sign stability for the headline family-TV-above-null direction across all 19 non-degenerate configurations. Family-TV does not affect family construction, which uses block co-visitation. The typed-state Family-TV diagnostic does use role labels, but the OAT sweep shows its sign and qualitative conclusions are stable under all non-degenerate role-threshold variants.

C.2 Representation Ablation: Per-(Pipeline×Model)

The full representation ablation discussed in §4.3 is reported here. Pipelines are mutual-kNN graphs with the same BCC decomposition and vary only the slice representation: cosine on the same activation keys; PCA-cosine at $d \in \{2, 10, 50\}$; and `all-MiniLM-L6-v2` sentence embeddings on 256-token slice text. "collapsed" indicates a graph too degenerate for a useful BCC decomposition. We interpret the MiniLM failure as a same-problem saturation effect: within one problem, slice text embeddings preserve topical relatedness but compress many slices into a narrow similarity range, leaving too little within-problem variance for mutual-kNN to separate local trunks, forks, and re-joins.

Table 8: Representation ablation, full per-(pipeline×model) family geometry breakdown: multi-fam % / mean correct families per cell on the same graphs as Table 1 of the main text.

Pipeline	R1-8B	Think	Inst
Jaccard kNN (ours)	$\mathbf{89.3/5.14}$	$\mathbf{86.2/4.61}$	$\mathbf{81.1/4.21}$
Cosine kNN	$88.1/4.80$	$85.5/4.36$	$79.6/3.87$
Overlap kNN	$87.7/4.77$	$85.5/4.36$	$79.6/3.87$
PCA-50 + cos	$81.4/3.38$	$81.4/3.33$	$73.8/3.00$
PCA-10 + cos	$84.9/3.45$	$81.6/3.40$	$73.5/2.96$
PCA-2 + cos	$83.3/3.65$	$82.1/3.76$	$74.1/3.52$
MiniLM slice	*collapsed* (n=5)	*collapsed* (n=5)	*collapsed* (n=11)

C.3 Fixed-Scaffold Split-Half Stability

As a lightweight held-out-style check, we split each cell's runs into two deterministic $32/32$ halves after constructing the canonical SliceGraph and non-trivial block scaffold. Within each half we induce the run-level co-visitation graph on normalized-correct runs only and rerun the same Louvain family detector. This does not yet prove that a graph built from one half can project every held-out run into a matched family, but it tests whether the observed family multiplicity depends on a small set of runs that vanish under resampling.

Across 779 non-code primary cells with at least two correct runs in both halves, both halves are multi-family in 97.2% of cells (problem-clustered bootstrap 95% CI $[96.0, 98.2]\%$). The full-cell multi-family rate on this split-valid subset is 99.7%. The mean absolute difference in half-level family count is 1.12 families (median 1), and the two half-level pairwise family-isomer rates are 75.1% and 75.7%, with median absolute pair-rate difference 5.1 points.

We treat this as a lightweight stability check. The stricter held-out validation below completes the picture.

C.4 Held-Out Atlas Validation

To move beyond a fixed-scaffold stability check, we perform a fully held-out validation: (i) build a SliceGraph from a random 32-run *train half*; (ii) for each held-out slice, compute nearest neighbours against train-half slices only using its sparse activation keys (without test-test edges or the full 64-run graph topology) and assign it to that neighbour's primary non-trivial block; (iii) compute covisitation families on the held-out correct runs using the train-half block scaffold; and (iv) measure whether the held-out partition is multi-family and how it relates to the full-sample partition.

We run $R{=}5$ random splits per cell across the 834 isomer-eligible primary cells (§B.1). Block coverage from test onto train blocks is $\mathbf{95.5}\%$, confirming that the scaffold generalizes. Held-out cells are multi-family in $\mathbf{91.0}\%$ of cases (vs. 97.8% on the full 64-run sample), with mean held-out pairwise isomer rate $\mathbf{60.1}\%$ (vs. 77.8% for the matched full-sample partition). NMI between full-sample and held-out family labels is moderate (0.24 median), indicating that exact family identities are not stable under a half-sample scaffold. We therefore use this held-out check to support multi-family presence and block coverage, not one-to-one family recovery.

C.5 Correct-Run Controlled Isomer Rate

Family count might be confounded by the number of correct runs: cells with higher accuracy have more correct runs and might discover more families simply due to sample size. We control for this by fixing the number of correct runs to $m \in \{4, 8, 16\}$, subsampling m correct runs per cell (with $R{=}10$ replicates) and recomputing the pairwise isomer rate on the same block scaffold.

Table 9: Pairwise isomer rate at fixed correct-run count m. Mean over $R{=}10$ replicates per cell, pooled across three primary models. Multi-family structure persists even at $m{=}4$.

	$m{=}4$	$m{=}8$	$m{=}16$
Mean isomer rate	58.2%	69.6%	75.2%
Multi-family %	82.1%	99.7%	100%
Mean families	2.3	3.3	4.4

Even at $m{=}4$, the multi-family rate is 82.1% and the pairwise isomer rate is 58.2%, still substantial. At $m{=}8$ the rate already recovers to $\sim$70%, and at $m{=}16$ it nearly matches the full-sample 76.6% headline. The claim is therefore not an artefact of high correct-run counts inflating opportunity.

C.6 Family Construction Sensitivity

We sweep the two main Louvain hyperparameters—resolution $\rho \in \{0.5, 1.0, 1.5, 2.0\}$ and weighted-Jaccard edge threshold $\tau \in \{0.025, 0.05, 0.1\}$—across 834 primary cells.

Table 10: Pairwise isomer rate (full-cell block weights) under a resolution×threshold sweep on the correct-only covisitation graph. Canonical configuration: $\rho{=}1.0$, $\tau{=}0.05$.

Resolution	$\tau{=}0.025$	$\tau{=}0.05$	$\tau{=}0.1$
0.5	0.247	0.365	0.552
1.0	0.744	**0.766**	0.793
1.5	0.857	0.865	0.868
2.0	0.917	0.915	0.913

At $\rho{=}1.0$ the isomer rate is stable across thresholds (74–79%), and the multi-family rate exceeds 96% for all three threshold values. Lower resolution ($\rho{=}0.5$) under-partitions and higher resolution ($\rho{\geq}1.5$) over-partitions, but the multi-family presence is robust across $\rho{\geq}1.0$. Replacing full-cell block coverage weights with correct-only weights changes the canonical isomer rate by only 3.4 percentage points ($76.6\% \rightarrow 73.2\%$), confirming that incorrect runs' contribution to rare-block upweighting is not the driver.

Single-axis ranges. Table 11 reports the range of the pairwise isomer rate when one construction axis is varied at a time with all other axes at the reference configuration.

Table 11: Pairwise structural-isomer rate when one axis is varied at a time (reference configuration otherwise).

Varied axis	Pairwise isomer rate
Edge threshold τ=0.025–0.10	74.4–79.3%
Rare-block weighting	62.0–76.6%
Louvain resolution ρ=0.5–2.0	36.5–91.5%
Slice granularity (`sep_up`)	75.2–76.6%
Neighbourhood size k=4–12	47.6–82.7%

Fixed-scale pair persistence. To separate the recurrence of separation from the stability of exact membership, we hold slice granularity, k, mutuality, and ρ=1 fixed and re-partition each cell under nine configurations that vary the edge threshold, rare-block weighting, and Louvain seed. For each correct-run pair we record the fraction of configurations in which it is separated. 52.6% [51.5, 53.7] of pairs are separated in at least 80% of the nine configurations; the distribution is strongly bimodal, with 48.4% of pairs separated in 90–100% of configurations and 9.7% separated in at most 20%. When the envelope is broadened to the 32 configurations of the audit in §C.1, which also change graph scale and mutuality, the corresponding consensus rate is 14.7%.

C.7 Isomer-Discovery Curves under Run Subsampling

We subsample the existing N=64 caches at $N \in \{8, 16, 32, 64\}$ (R=5 replicates, uniform without replacement) and recompute family detection and core masks at canonical hyperparameters.

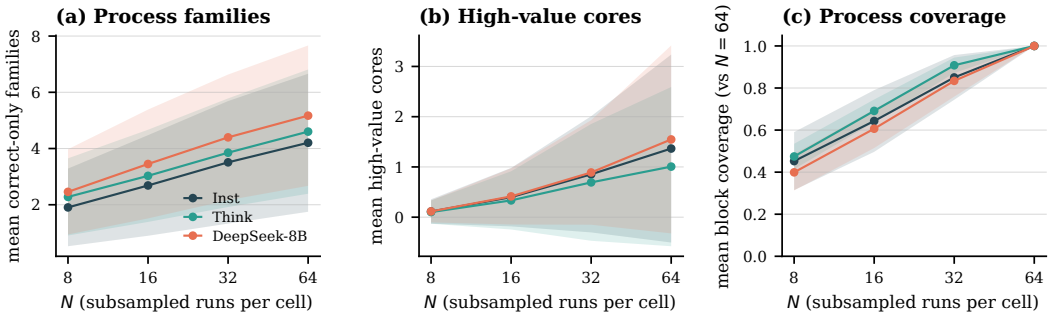


Figure 4: Isomer-discovery curves under run subsampling. (a) Correct-only family count, (b) high-value-core components, (c) block coverage relative to N=64. Block coverage saturates around $N \approx 32$; family and core counts keep growing as rare-block neighbourhoods enter the sample.

D Process-Isomer Geometry and Dynamics

The reward field is used in this paper only as a *label-seeded descriptive overlay*. A cell enters the reward-evaluable subset only when the leave-one-out seed is stable enough to define such an overlay; this is a construction requirement, not a classification objective.

Reward-field propagation details. For a cell with n graph-valid runs and block visitor set $\mathcal{R}_b$, the global seed used for high-value cores is

$$v_b^{(0)} = \mathbf{1}[3 \le n_b \le n-3] \left(\frac{1}{n_b} \sum_{r \in \mathcal{R}_b} y_r - \frac{1}{n} \sum_r y_r \right) \left(\frac{n_b}{n} \right)^{1/2}, \qquad n_b = |\mathcal{R}_b|. \tag{6}$$

The run-level leave-one-out variant used for per-run overlays removes that run from both the block and cell denominators. The diffusion graph has non-trivial blocks as nodes; K_2 bridges are not

Table 12: Full model×dataset route multiplicity including replication models. Entries: % cells with ≥2 correct-only families / mean correct families. Top: primary corpus; bottom: replication models. Llama-8B covers AIME24, AIME25, and GPQA only.

Model	AIME24	AIME25	BRUMO25	GPQA	HMMT25	All
R1-8B	86.7 / 4.87	86.7 / 4.80	86.7 / 4.57	92.4 / 5.50	76.7 / 3.97	89.3 / 5.14
Think	93.3 / 5.07	86.7 / 5.10	96.7 / 4.77	85.9 / 4.60	70.0 / 3.57	86.2 / 4.61
Inst	83.3 / 4.20	70.0 / 3.43	76.7 / 3.70	87.9 / 4.72	50.0 / 2.17	81.1 / 4.21
Llama-8B	76.7 / 4.80	53.3 / 3.10		82.3 / 3.98		78.3 / 3.98
Qwen3-32B	93.3 / 5.17	90.0 / 4.93	96.7 / 4.80	88.4 / 4.56	86.7 / 4.17	89.6 / 4.64
Q2.5-72B	46.7 / 1.93	36.7 / 1.33	46.7 / 2.40	86.4 / 4.18	13.3 / 0.47	67.3 / 3.18

diffusion nodes. Let E_{BC} connect two non-trivial blocks that share an articulation point in the block-cut tree, and let E_{temp} connect consecutive distinct primary blocks in a compacted run path. Temporal edges are symmetrized, multiplicities are binarized, and the raw adjacency is

$$A^{\mathrm{raw}}_{ij} = \mathbf{1}[i=j] + \mathbf{1}[(i,j) \in E_{\mathrm{BC}} \cup E_{\mathrm{temp}}], \qquad P_{ij} = \frac{A^{\mathrm{raw}}_{ij}}{\sum_k A^{\mathrm{raw}}_{ik}}. \tag{7}$$

Thus block-cut and temporal edges have equal weight ($\lambda = 1$) after binarization, and the self-loop weight is 1. In all experiments we use α=0.65 for 24 steps,

$$\mathbf{v}^{(t+1)} = \alpha\, \mathbf{v}^{(0)} + (1-\alpha)\, \mathbf{P}\mathbf{v}^{(t)}. \tag{8}$$

We keep $\mathbf{v}^{(24)}$ as the field used throughout §4.6; the high-value core is the top quartile of its strictly positive support. After 24 iterations, each propagated field—pooled or family-conditioned—is independently rescaled by its own ℓ_∞ norm. Core masks, connected components, and specialization statistics are invariant to positive rescaling; field-sharpness ratios (Equation 10) should be read under this per-field normalization convention.

Alignment and non-redundancy definitions. For a cell with high-value-core components $C_1, \ldots, C_K$ and a correct-only family f, let B_f be the union of non-trivial blocks visited by runs in f and $m_f(k) = |B_f \cap C_k|/|C_k|$. For $K \geq 2$ we normalize each non-zero row to $p_f(k) = m_f(k)/\sum_j m_f(j)$ and define

$$\mathrm{Spec} = 1 - \frac{1}{|F_+|} \sum_{f \in F_+} \frac{H(p_f)}{\log K}, \tag{9}$$

where F_+ are families with non-zero core footprint. The reported 0.59 is averaged on the 324 multi-core reward-evaluable cells. For a family-conditioned field $\mathbf{v}_f$, computed by setting correctness labels outside family f to zero before applying Equations 6–8, field sharpness is

$$\mathrm{Sharp}(f) = \frac{\max_i v^{(24)}_{f,i}}{\max_i v^{(24)}_i}. \tag{10}$$

Coverage loss uses the most coverage-critical family in a cell:

$$\Delta_{\max} = \max_f \frac{|(\cup_g B_g) \setminus (\cup_{g \neq f} B_g)|}{|\cup_g B_g|}. \tag{11}$$

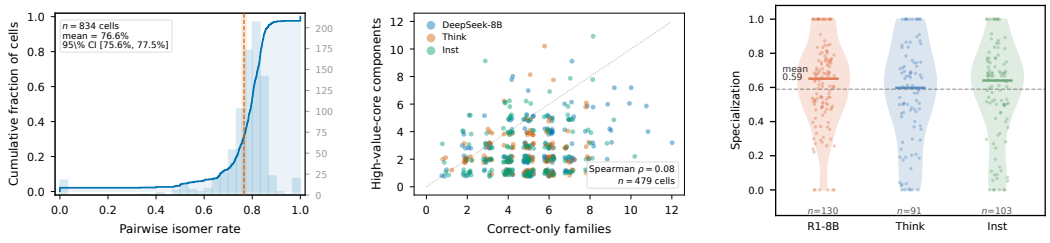
The 94.5% headline is $\Pr(\Delta_{\max} > 0.10)$ over the 816 canonical cells with at least two correct-only families; it measures the most coverage-critical family per cell, not every family.

Core divergence. For reward-evaluable cells, two correct runs are *core-divergent* if their non-empty sets of visited high-value-core components are disjoint. This relation characterizes divergence in the overlay rather than defining the route partition; it is evaluated only for runs with at least one core visit, and it adds +1.3 pp to the family-isomer rate on the cells where both are defined.

The pairwise family-isomer rate is the primary Definition 1 rate: it is computed within each problem-model cell that has at least two normalized-correct runs, as the share of unordered correct-run pairs assigned to different correct-only route families, then averaged across cells. Core divergence is reported separately as an overlay refinement (§4.6).

Table 13: Multi-modal reward geometry (non-code). Sections A–C use the three primary models; section D adds the Llama cross-architecture replication.

Statistic	Value
A. High-value core disconnectedness (n=479)	
% cells with ≥ 2 core components	67.6%
Mean / median # core components	2.61/2
Size ratio: 2nd / 1st component	0.54
B. Correct-only families (n=954)	
% cells with > 1 correct family	85.5%
Mean correct families / cell	4.66
Pairwise family-isomer rate (n=834 cells)	mean 76.6%, median 79.6%
Problem-clustered bootstrap CI for mean	[75.6, 77.5]%
Split-half both-halves multi-family (n=779 cells)	97.2%
Length-variance ratio	0.460
Modularity null z (median)	35.54
C. Alignment and non-redundancy (natural denominators)	
Family–core footprint specialization (multi-core cells, n=324)	0.59
Family field-max / pooled-global field-max (2,094 family fields)	1.55×
Jaccard(family core, global core)	0.17
Spearman(#components, #families)	0.08
% multi-family cells where most critical family removal loses $> 10\%$	94.5%
Core-divergence increment over family-isomer rate	+1.3 pp
D. Llama cross-architecture replication	
% with ≥ 2 core components (n=194)	78.4%
% with > 1 correct family (n=258)	78.3%
Family–core footprint specialization (multi-core)	0.60
Family field-max / pooled-global field-max	1.46×
Length-variance ratio	0.469



(a) Isomer rate ECDF. (b) Families vs. cores. (c) Core specialization.

Figure 5: **Population geometry (primary corpus).** (a) ECDF of pairwise family-isomer rates over 834 cells under the reference configuration; vertical lines mark the bootstrap 95% CI. (b) Correct-family count vs. high-value-core components on 479 reward-evaluable cells (Spearman ρ=0.08) and (c) family–core specialization on 324 multi-core cells (bars = medians; pooled mean = 0.59) are descriptors of the secondary reward overlay (§4.6 of the main text).

D.1 Per-Model Block-Type Localisation

The pooled reward-field statistics collapse a model-dependent block-type localisation that is itself informative. Inst concentrates at SHARED_TRUNK; R1-8B, Think, and Llama-8B concentrate at ANSWER_BASIN. All four therefore exhibit a localized high-value region; they differ only in which structural niche hosts it.

Table 14: Per-model mean propagated reward field by block-type role on the canonical reward-field corpus. Bold cells indicate the dominant block-type per model.

Model	shared_trunk	answer_basin	weak_basin	intermediate	decision_point
Inst	**0.138**	0.001	0.016	0.005	0.003
R1-8B	0.062	**0.252**	0.062	0.029	0.023
Think	0.064	**0.236**	0.046	0.030	0.023
Llama-8B	0.031	**0.123**	0.044	0.006	0.006

D.2 Family-Specific Navigation Diagnostics

The typed-state transition model is used to compare how different correct-only families traverse the same atlas. The population-level dynamics headline in the atlas version is family-specific kernel divergence rather than an auxiliary task objective.

For each run we map the primary non-trivial block at slice t to $z_{r,t} = (\mathrm{role}(b^\star_{r,t}), \mathrm{posbin}(\tau_t), \mathrm{core}(b^\star_{r,t}))$, compact consecutive duplicates, and append absorbing correct/wrong terminals. Transition counts C are Laplace-smoothed into

$$P_{ij} = \frac{C_{ij} + \alpha_{\mathrm{smooth}}}{\sum_{j'} C_{ij'} + \alpha_{\mathrm{smooth}}|\mathcal{S}|}, \qquad \alpha_{\mathrm{smooth}} = 0.5. \tag{12}$$

Family-TV averages row-wise kernel divergence,

$$\mathrm{TV}(P_f, P_{f'}) = \frac{1}{2|\mathcal{S}|} \sum_{i \in \mathcal{S}} \|P_f(i,\cdot) - P_{f'}(i,\cdot)\|_1. \tag{13}$$

In the released implementation, the averaging set $\mathcal{S}$ includes the two absorbing terminal rows ($\mathrm{EOS}_{\mathrm{correct}}$, $\mathrm{EOS}_{\mathrm{wrong}}$); because both are forced absorbing with identical rows across families, they contribute zero TV. Removing them rescales the denominator but does not affect the matched-null comparison reported in §4.6. The pooled-kernel committor solves $(I - Q)q = r_A$, where Q is the transient-to-transient block and r_A is the transition vector into $\mathrm{EOS}_{\mathrm{correct}}$. The three-step escape hazard is $h_3 = \sum_{i \in \mathcal{C}} \hat{\pi}(i) \sum_{j \notin \mathcal{C}} P^3_{ij}$; return and MFPT readouts use the same smoothed kernel and are retained as case-level diagnostics.

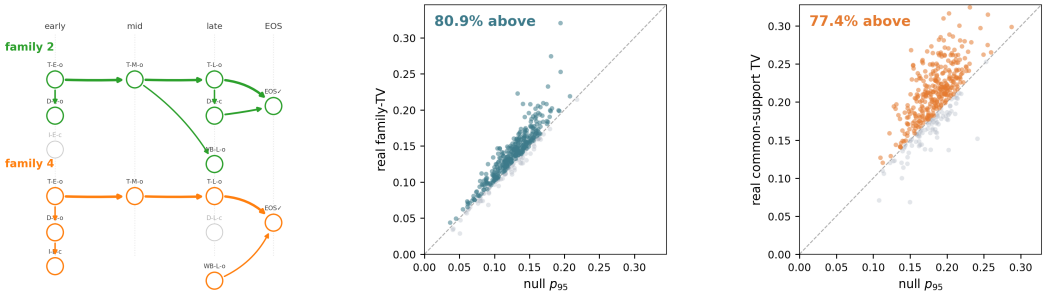


Figure 6: **Family-specific typed-state dynamics.** (a) Two families share a typed-state vocabulary but place different transition mass. (b, c) Real family-TV and common-support TV versus size-preserving label-shuffle p_{95}; points above the diagonal exceed the matched null.

Table 15: Family-specific navigation diagnostics on the core-eligible multi-family subset (n=403).

Statistic	Value
Median family pairwise kernel TV	0.137
% cells with real TV > null p_{95}	80.9%
Median null z	3.14
Corr.(family-TV, static specialization)	−0.07

Table 16: Common-support family-TV diagnostic. TV is averaged only over typed-state rows visited by both families in a pair; the null shuffles run-to-family assignments while preserving family sizes.

Statistic	Value
Eligible cells	403
Median common-support family TV	0.206
Median common-support null p_{95}	0.183
% cells with common-support TV $>$ null p_{95}	77.4%
Median common-support null z	3.45
Median shared typed states per family pair	7.0

The pooled-kernel committor, MFPT to core, escape hazard, and post-escape return remain useful *case-level atlas readouts* for understanding how one family navigates relative to another. In this atlas version they are not treated as standalone population claims.

D.3 Typed-State Sensitivity and Scope

Family-TV is the only population-level dynamics statistic we headline, because it is the most stable to mild changes in smoothing and role thresholds. The companion pooled committor, MFPT, escape, and return diagnostics remain informative for inspecting one fixed atlas, but their absolute numeric values are more sensitive to role-threshold choices and sparse-transition smoothing. We therefore use them as descriptive overlays on the process atlas rather than as independent paper claims.

D.4 Null-Control Ledger

Table 17 maps each main-text claim to the corresponding weak reading and diagnostic used to control it.

Table 17: Claim-to-control ledger for the main atlas readouts.

Claim	Weak reading	Control / diagnostic
Families are route units	arbitrary graph clusters	human audits, representation ablation, modularity null
Structural isomers are common	few-run or configuration artifact	non-singleton rate, fixed-scale persistence, split-half, held-out atlas
High-value cores are multi-component	heatmap fragmentation	core-size ratio, label-permutation caveat (descriptive)
Family dynamics differ	support mismatch only	family-label null, common-support TV
Route-collapsed consensus is less reliable	vote-share confound	vote-share strata, matched within-problem comparison

The paper uses several nulls, each aimed at a different possible artifact. We use five null families, each targeting a different possible confound:

1. **Degree-preserving graph rewire** — preserves the degree sequence but destroys local adjacency; used for the family modularity headline (median z=35.54).
2. **Block-type-preserving rewire** — preserves role counts and type marginals; used as a structural-sensitivity stress test (effect-drop only).
3. **Family-label shuffle** — preserves family sizes and typed-state support but destroys route-specific labels; the exported population null for family-TV (80.9% above p_{95}, median z=3.14).
4. **Temporal-order shuffle** — preserves visited typed states but destroys transition order; used for kernel, escape, return, and MFPT stress tests (effect-drop only).
5. **Label permutation** — preserves graph topology, family partition, and per-cell label count but randomises correctness association; the reward-core stress test (Table 18).

Items 1 and 3 generate the exported headline nulls; items 2 and 4 are validity-ladder stress tests and should not be read as independent population claims.

Table 18: Label-permutation reward-field stress test. Permutations preserve the graph topology, family partition, and the number of correct labels in each reward-evaluable cell, but randomize which runs are correct. The permutation range reports shuffle-level p_5–p_{95} across 100 shuffles.

Readout	Real	Label-perm. mean [p_5,p_{95}]	z
Multi-core rate	67.6%	85.4% [83.3,88.1]%	−11.44
Family–core footprint specialization, real multi-core cells	0.59	0.74 [0.72,0.76]	−11.87
Family field-max / pooled-global field-max, field-weighted	1.55×	1.55× [1.48,1.70]	0.08
Jaccard(family core, global core), field-weighted	0.17	0.10 [0.097,0.105]	29.36

Label permutations tend to scatter positive seed mass, which can create more disconnected random cores and higher apparent specialization than the real labels. We therefore do not use this stress test as a one-sided null for multi-core counts. Its main diagnostic value is that real family-conditioned cores overlap the real global core far more than permuted labels do, while field sharpness alone is comparable to the permutation range and is interpreted only together with the family/core geometry above.

D.5 Family-TV Null Construction

For each core-eligible multi-family cell, the family-TV null keeps family sizes and typed-state support fixed while shuffling family labels across trajectories before re-estimating per-family kernels. This preserves the atlas and its coarse traffic levels while erasing route-specific navigation structure. The main-text 80.9%-above-null and median-z=3.14 headline is measured against this null.

Common-support TV. A stricter diagnostic averages TV only over typed states visited by both families in a pair. This tests whether families move differently even where they share support, rather than only because they visit different states. Table 16 reports the matched family-size shuffle null and the shared-support denominator summary.

D.6 Worked Route Report on GPQA-71

As a worked example we pick GPQA-71 under **Inst** (Qwen3-4B-Instruct-2507): seven correct-only route families covering 38 non-trivial blocks across four disconnected high-value-core components. Removing the single most coverage-critical family loses 55.3% of total block coverage, and the loss-if-removed shares sum to 92.1%, confirming that route coverage is family-unique rather than redundantly shared. The four core components are occupied by different families (F6→core 0, F0→core 1, F2/F5→core 2/3), while F1, F3, F4 are off-core under the top-quartile mask. Per-family coverage plots and the full summary card are included in the supplementary material.

D.7 Route-Collapsed Consensus and Family-Aware Voting

This section gives the protocol behind §4.5.

Label-free families. For each cell we fit a second family partition F_{all} on *all* graph-valid runs, using the same block co-visitation weighted Jaccard, threshold, weighting, and Louvain settings as the reference configuration but without any correctness filtering. Correctness labels are used only to score the resulting flags and aggregation rules.

Cohorts. Cohorts of $N \in \{8, 16, 32, 64\}$ runs are drawn from the canonical N=64 caches. Block regions and family assignments for the subsampled runs come from the amortized unlabeled 64-run atlas, so smaller cohorts establish retrospective small-cohort analysis, not online compute savings.

Aggregation rules. Self-consistency (SC) returns the most frequent normalized answer. *One family, one vote* lets each family of F_{all} cast one vote for its majority answer; $\sqrt{\textit{size}}$ *weight* weights each

family's vote by the square root of its size; *family support* scores each answer by the number of families supporting it. Ties follow SC's tie rule. Table 3 in the main text reports accuracies; the family-aware rules trail SC by 0.19–0.85 points at every N, matched within-problem comparisons are mostly null, and family support does not improve global AUROC.

Route-collapsed consensus. A cohort's SC majority is *route-collapsed* when all of its supporting runs belong to one family of F_{all}. At $N{=}8$ the flag marks 7.8% of cohorts; flagged cohorts are correct in 33.6% of cases, the remaining cohorts in 69.2%, and all cohorts in 66.4%. At $N{=}16$ and $N{=}32$, within vote-share strata that contain both single-family and multi-family majorities, the multi-family majorities are 14.6–44.4 percentage points more accurate. Because the comparison is across problems, part of the gap may reflect problem difficulty; the matched within-problem comparisons above are mostly null, so the demonstrated signal is problem-level auditing of structurally narrow consensus rather than within-problem answer selection.

E Annotation Protocols and Family Robustness

This section covers family-detection robustness and the two human-annotation tasks used in §4.2. Task A audits BCC shared-state semantics; Task B audits route-family same-strategy semantics.

Table 19: Compact summary of annotation and family-validation statistics, split by task. Top block: Task B (route families); bottom block: Task A (BCC shared-state).

Statistic	Value
Task B (route families)	
Within-family decisive agreement (Annotator A)	94.4%
Self-consistency κ (Annotator A)	0.65
Across-family cross-annotator κ (A $\times$ B)	**0.737**
Decisive exact agreement, family pack (A $\times$ B, $n{=}55$)	81.8%
Task A (BCC shared-state)	
BCC agreement vs. ground-truth (Annotator B, $n{=}72$)	**90.3%**
BCC agreement vs. ground-truth (Annotator A, $n{=}72$)	88.9%
BCC-pack cross-annotator agreement (A $\times$ B, both decisive, $n{=}71$)	**97.2%**
Matched non-articulation specificity (Annotator B)	36/36

E.1 Family Detection Robustness

The default 85.5% multi-family rate is not Louvain-specific: greedy modularity reproduces the same direction and magnitude, while a degree-preserving modularity null is decisively weaker (median $z{=}35.54$). Label propagation is included only for transparency and is not preferred because it collapses weighted-Jaccard family graphs pathologically.

E.2 Annotation Protocol

Annotators A and B are two non-author external annotators with mathematical reasoning and LLM-trace experience. They are blind both to the project hypothesis and to within/across pair labels, and cross-label both the BCC pack (Task A) and the route-family pack (Task B) under matched strict rubrics.

Pair construction. Task A (BCC): for each BCC block, a within-BCC pair is formed by selecting two runs from different trajectories with the smallest positional gap; a matched across-BCC control with the same correctness label and the smallest absolute position offset is paired. For articulation transitions, the first positive articulation-boundary crossing per trajectory is matched with a phase-matched non-articulation control. Task B (families): 8 cases are selected across 3 models and 3 datasets, each with ≥ 4 correct-only families and ≥ 5 decision forks; 2 runs per family are sampled from normalized-correct runs only, balanced by length and model/dataset via stratified sampling, yielding 46 within-family + 31 across-family pairs from 102 runs.

Blinding and randomisation. The self-consistency re-review uses a strictly blinded prompt template that hides: (i) within-vs-across family labels, (ii) original first-pass judgements, (iii) family IDs. Pair order is randomised. Pair identity is tracked by a deterministic 12-character hash of (case_id, sorted run IDs). The cross-annotator wave uses the same blinded pack in the same randomised order, in an independent session with no access to the primary annotator's responses.

Response schemas. Task A: `annot_same_reasoning_state` $\in$ {yes, no, uncertain}, `annot_confidence` $\in$ 1–5, plus free-text `annot_notes`. Task B: `same_strategy` $\in$ {yes, no, uncertain}, `confidence` $\in$ 1–5, `justification` (one sentence).

Uncertainty handling. All tasks use a three-value scheme (yes/no/uncertain). The instruction states: "Use `uncertain` sparingly—only when two runs share a framework but execute it in materially different ways." We report two denominators: *raw* (uncertain counts as negative) and *decisive* (uncertain excluded from both numerator and denominator). The decisive rate is the primary claim metric; raw counts or uncertain counts are reported where they affect interpretation.

Cohen's κ calculation. We use unweighted Cohen's κ with the label set {yes, no, uncertain}. For the across-family sub-task, labels are inverted (`different_strategy=yes` $\rightarrow$ `same_strategy=no`) before computing κ. Because κ is computed on the three-value label set, its denominator is the set of pairs for which both judgements are available, whereas the decisive rates above exclude `uncertain` judgements. Key values: Annotator A self-consistency κ=0.65 (76 doubly-judged pairs), across-family-only κ=0.80; cross-annotator across-family κ=0.737 (30 across-family pairs judged by both annotators).

E.3 Task A: BCC Shared-State Audit

Protocol. The core BCC pack contains 72 slice-text pairs: 36 within-BCC pairs and 36 length-matched across-BCC controls. Both annotators judge whether the two slices represent the same reasoning state. Separately, we construct an articulation transition pack of 72 adjacent-slice comparisons (36 true articulation crossings and 36 matched non-articulation controls); this auxiliary pack is used as a precision-side control rather than as a second headline task.

Cross-annotator BCC agreement. On the 72-pair BCC slice-text pack, Annotator B agrees with the BCC-derived labels on $\mathbf{90.3}\%$ of pairs and Annotator A reaches 88.9%. The two annotators agree with each other on 69 of the 71 pairs on which both are decisive ($\mathbf{97.2}\%$), i.e. two disagreements. This is the strongest evidence that the BCC/shared-state interpretation is independently recoverable.

E.4 Task B: Route-Family Same-Strategy Audit

Protocol. We select 8 cases across the 3 primary models and three datasets (AIME24, HMMT25, and GPQA), requiring at least 4 correct-only families, at least 5 decision forks, and problem accuracy between 40–80%. From these cases we sample 102 runs, yielding a final judged pack of 77 blinded run pairs: 46 within-family and 31 across-family. Input CoTs are truncated to 15,000 characters via a front/middle/tail (5k/5k/5k) scheme with omissions marked explicitly.

Cross-annotator family agreement. The numbers referenced in the main text are: (i) Annotator A's within-family decisive agreement rate of 94.4% ($34/36$ decisive pairs judged same-strategy); (ii) cross-annotator κ=0.737 on the across-family slice (three-value labels on the n=30 across-family pairs judged by both annotators; raw agreement 90.0%); and (iii) 81.8% exact agreement on the subset where both annotators give a decisive judgement (n=55).

Across-family differentiation rate. Of 31 across-family pairs, Annotator A gives 10 decisive "different strategy" judgements, 11 decisive "same strategy," and 10 uncertain. The decisive differentiation rate is therefore $10/21$=47.6%; counting the ten uncertain judgements as not establishing a difference gives $10/31$=32.3%. Both are conditional rates on eight deliberately selected family-rich cases, not corpus-level prevalences, and we do not multiply them by the structural isomer rate. Partial separation is consistent with SliceGraph families being defined by co-visitation of non-trivial blocks (an activation-defined, segment-level signal) that can separate runs which share a human-named

strategy but differ in intermediate operations; some boundaries may also reflect representation or partitioning effects, which this audit cannot distinguish. Under a matched strict rubric, Annotator B's family-level yes-rate is within 4 pp of Annotator A's, so the family claim is not a single-annotator artifact.

E.5 Qualitative Annotation Cases

Table 20 summarises four representative cases used for qualitative inspection; full anonymised verbatim excerpts are included in the supplementary material as `annotation_cases/`.

Table 20: Qualitative annotation cases. Confidence is on a 1–5 scale.

Case	Task	Judgement	Conf.	Diagnostic point
A1	BCC within/across	same vs. different state	4/5	Answer-finalisation pair vs. quadratic-solving control (AIME24)
A2	Articulation transition	strategy shift	5/5	Divisor enumeration → prime-factorisation pivot (HMMT25)
B1	Within-family	same strategy	5/5	Both runs set up binary-expansion identity identically (HMMT25)
B2	Across-family	process-isomer pair	5/5	Numerical-first vs. theory-first route to same answer (HMMT25)

F Cross-Model Replications

F.1 Cross-Scale Replication on Qwen2.5-Math-72B-Instruct

We replicate the full canonical pipeline—including the reward field, multi-core decomposition, family-core specialization, and family-TV—on Qwen2.5-Math-72B-Instruct (20,352 sampled runs, 20,196 graph-valid rows; 318 non-code problems on AIME24/25, BRUMO25, GPQA, HMMT25) under the identical configuration (`sep_up`=8, k=6, mutual-kNN, σ=0.35, M=2,600; reward α=0.65, `core_q`=0.75).

Process-family heterogeneity. Under canonical Louvain, 72B retains multi-family structure but attenuates on the math benchmarks (67.3% of cells overall; Table 12). On GPQA the multi-correct-family rate is **86.4**%, essentially matching the 3-model 85.5% headline. One candidate explanation is that this math-specialized model solves many math problems near-deterministically, but we do not report per-model accuracy and therefore do not attribute the attenuation to accuracy saturation; on HMMT25 the mean number of correct families is below one, so part of the attenuation reflects cells without a valid correct family.

Reward-field and multi-core structure. The multi-core rate drops to 42.9% (vs. 67.6% primary). On the 81 cells that *do* exhibit multi-core structure, family-core specialization is higher (0.738 vs. 0.59); as this readout is descriptive (its label-permutation mean on the primary corpus is 0.74, Table 18), we do not read it as sharper alignment. Family-conditioned field-max ratios (1.57×) and mean family-TV (0.117) match the primary pattern (1.55×, 0.135).

Table 21: 72B replication summary. "Primary" is the pooled 3-model value cited in the main text; "72B" is the matched replication on the Qwen2.5-Math-72B-Instruct corpus.

Main construct	Primary (3-model)	72B	Conclusion
multi-core rate	67.6%	42.9%	attenuated
family×core spec. (multi-core cells)	0.59	0.738	higher when multi-core
family / pooled-global field-max	1.55×	1.57×	matched
multi-correct-family rate	85.5%	67.3% (mean 3.18/prob)	attenuated
length-variance ratio	0.460	0.484	not a length artifact
mean family-TV	0.135	0.117	matched

Dataset-level values are included in Table 12. The overall pattern is that the 72B replication attenuates multi-core and multi-family rates while retaining the qualitative atlas structure on GPQA. Because model generation, training data, architecture, and task performance all differ, this comparison does not isolate a model-size effect or establish its cause.

F.2 Cross-Scale Replication on Qwen3-32B

We replicate the full canonical pipeline on Qwen3-32B (20,352 sampled runs; 318 non-code problems on AIME24/25, BRUMO25, GPQA, HMMT25) under the identical configuration. After degenerate-correctness filtering, 147 reward-evaluable cells remain.

Key result. Qwen3-32B closely matches the primary 3-model atlas picture and shows higher measured multiplicity than the 72B replication on multi-family and multi-core rates. The multi-correct-family rate (89.6%) exceeds the primary headline (85.5%); multi-core rate (75.5%) matches Llama-8B (78.4%); family-TV (0.143) and family/global field-max (1.60×) are numerically matched to the primary (0.135, 1.55×). Unlike 72B, Qwen3-32B retains the full multi-route structure on the math benchmarks.

Table 22: Qwen3-32B replication summary.

Main construct	Primary (3-model)	32B	Conclusion
multi-core rate	67.6%	75.5%	matched / above
family×core spec. (multi-core cells)	0.59	0.52	comparable
family / pooled-global field-max	1.55×	1.60×	matched
multi-correct-family rate	85.5%	89.6% (mean 4.64/prob)	matched / above
length-variance ratio	0.460	0.440	not a length artifact
mean family-TV	0.135	0.143	matched

F.3 Cross-Architecture Replication on DeepSeek-R1-Distill-Llama-8B

The Llama cross-architecture replication (16,512 runs across 258 problems on AIME24, AIME25, GPQA) is the strongest single piece of architecture transfer evidence. The same atlas picture reappears: multi-core reward overlays, several correct-only families per problem, and near-identical family-core specialization.

Table 23: Llama-8B replication summary. "Primary" is the pooled 3-model value cited in the main text; "Llama" is the matched replication on the Llama corpus.

Main construct	Primary (3-model)	Llama	Conclusion
multi-core rate	67.6%	78.4%	multi-core dominant
family×core spec.	0.59	0.597	numerically matched
family / pooled-global field-max	1.55×	1.46×	matches sharpness
multi-correct-family rate	85.5%	78.3% (mean 3.98/prob)	multi-family dominant
length-variance ratio	0.460	0.469	not a length artifact
mean family-TV	0.135	0.127	matched
family-TV $>$ null p_{95}	80.9%	78.3% (n=152)	matched

Llama also matches the block-type localisation pattern of the reward field: its high-value region is answer-basin dominated (Table 14), and its family-level partition remains process-like rather than length-clustered. Family-TV dynamics replicate as well: across 152 Llama cells with valid typed-state kernels, 78.3% exceed the family-label shuffle null's p_{95}, matching the primary rate of 80.9%. Taken together, the Llama replication shows that the atlas story is not tied to one architecture within the Qwen family.