HIDDEN APIS IN LANGUAGE MODELS: DISCOVERING REUSABLE CAUSAL INTERFACES FROM FORKED FUTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Identical language-model answers can arise from hidden states that support different future computations, so present-answer decodability leaves the organization of reusable state unresolved. We introduce forked futures: future operations are sampled only after a prefix state has formed, and the resulting response distributions define a future causal signature over that state. Shared, Local, Mixture, and Distributed interfaces then compete under prequential causal description length with matched capacity and future-signature fidelity. Shared achieves the lowest held-out description length on Qwen2.5-1.5B and Llama-3-8B, with Sharedness Gains of 0.216 and 0.294 nats, while preserving comparable or lower mean, worst-family, and tail distortion; the preference remains positive across a five-backbone sweep. Role-aligned transplantation gives Shared the strongest joint target-correctness, locality, and copy-preservation profile, and API-aligned paths mediate 0.749 of the target effect versus 0.150 for matched null paths. In blind model organisms with known routing, the procedure recovers 14/16 architectures. Together, these results identify a compact reusable causal interface through convergent evidence from description length, held-out generalization, transplantation, mediation, and ground-truth architecture recovery.

1 INTRODUCTION

Language models compress contextual information into hidden states that are repeatedly consumed by downstream computation. Mechanistic interpretability therefore requires more than asking what a state decodes at the present output. Two prefixes can produce the same current answer yet support different later comparisons, verifications, executions, or compositions, while surface-different prefixes can support the same future behavior (Belinkov, 2022; Elazar et al., 2021). A representation can thus be highly decodable without revealing whether its information is organized for one immediate prediction or for reuse by several later consumers. The central question here is architectural: which internal variables form a reusable interface, how that interface is organized, and how it can be distinguished from local, gated, or distributed alternatives when the present answer alone is uninformative?

We introduce *forked futures*, a counterfactual framework that defines hidden-state identity by what future computations the state can support. A prefix is processed and its hidden state is cached before any future-operation cue is revealed; only then are operations sampled from an operation bank $\mathcal{K}$. Their response distributions form a *future causal signature*, and states with similar signatures define an empirical causal quotient. Because the future cue arrives after state formation, the quotient is organized by reusable transition structure rather than by anticipation of a particular operation. We compare four organizations of that quotient: Shared uses one operation-agnostic interface, Local uses operation-specific interfaces, Mixture gates among partially shared components, and Distributed retains future-relevant information across several components. Prequential causal description length selects the organization that predicts held-out future signatures most economically at matched fidelity (Dawid, 1984; Blier & Ollivier, 2018); a positive Sharedness Gain means that distinctions learned for earlier consumers reduce the code length of later consumers.

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.

Architecture selection is followed by direct causal tests. Role-aligned transplantation measures target correctness, locality, and copy preservation; producer–consumer mediation tests whether the recovered routes carry effects during ordinary computation; representation controls compare against J-space, SAE/Transcoder, PCA, output-gradient, answer-token, and random subspaces; and blind model organisms provide known routing ground truth. On Qwen2.5-1.5B and Llama-3-8B, Shared has the shortest held-out causal description with Sharedness Gains of 0.216 and 0.294 nats while preserving comparable or lower mean, worst-family, and tail distortion. API-aligned paths mediate 0.749 of the target effect versus 0.150 for matched null paths, natural-family transfer preserves the architecture ordering, and blind recovery identifies $14/16$ model-organism architectures. The evidence therefore links three levels of the claim: future sufficiency, economy of reuse, and causal consumption by downstream computation.

Our contributions are:

1. **Future-defined causal equivalence:** forked futures identify hidden states by response distributions over operations sampled only after the state has formed.
2. **Architecture discovery:** Shared, Local, Mixture, and Distributed interfaces compete under prequential causal MDL with explicit future-signature fidelity.
3. **Mechanistic validation:** transplantation, locality, donor-copy, mediation, necessity–restoration, graph recovery, and matched representation controls connect future sufficiency to causal use.
4. **Falsification and generalization:** known-routing model organisms and held-out operations, compositions, generators, natural families, languages, and backbones test the recovered organization beyond its fitting conditions.

2 Related Work

Global-workspace theories describe cognition as interaction between specialized processors and a capacity-limited shared state that makes selected information broadly available (Baars, 1993; Dehaene et al., 1998; Dehaene & Naccache, 2001; Mashour et al., 2020). Machine-learning instantiations use shared workspaces to coordinate modular networks and exchange information among otherwise specialized components (Bengio, 2017; Goyal et al., 2021), while recent work on language models reports representations that are broadly accessible or verbalizable across tasks (Gurnee et al., 2026). These results motivate a concrete architectural question: whether repeated computation is supported by one common interface or by several local or distributed states. Forked futures makes this distinction testable by placing Shared, Local, Mixture, and Distributed organizations in the same held-out competition and asking whether a common code earns an economy when multiple later operations consume the same prefix state. This shifts the workspace question from broad accessibility to a quantitative comparison of alternative internal organizations under the same future-behavior target.

Causal abstraction formalizes when a low-level system implements a higher-level causal model (Massidda et al., 2023; Geiger et al., 2025; 2021; Beckers & Halpern, 2019). Interchange interventions, path patching, Patchscopes, and automated circuit discovery extend this perspective by asking whether internal variables or paths causally support model behavior (Ghandeharioun et al., 2024; Conmy et al., 2023; Goldowsky-Dill et al., 2023). Most such analyses begin from a selected behavior, representation, or circuit and test its causal role. Our framework instead defines the object through a distribution over future consumers before testing its mechanism. The future causal signature specifies which distinctions must survive across several operations; role-aligned transplantation tests whether the same recovered code can be consumed by multiple readers, and path-blocked mediation identifies the producer–consumer routes through which that code contributes during the model's ordinary forward computation.

Representation probing and steering provide a complementary view of internal structure. Linear probes, representation engineering, activation addition, task vectors, sparse autoencoders, transcoders, PCA, and ICA reveal low-dimensional or sparse structure and can support effective behavioral control (Zou et al., 2023; Park et al., 2023; Turner et al., 2023; Hendel et al., 2023; Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2025; Dunefsky et al., 2024; Jolliffe, 2025; Hyvärinen & Oja, 2000). These methods motivate strong baselines because a compact or steerable subspace need not be a reusable causal interface. We therefore compare the recovered interface with

J-space, SAE/Transcoder, PCA, output-gradient, answer-token, and random subspaces on future-signature fidelity, description length, role-aligned transplantation, locality, selective necessity, and mediation. J-space provides a strong reusable-state reference, while the remaining controls test whether the observed effect can be explained by dimensionality, reconstruction, output sensitivity, or intervention load alone.

Minimum-description-length methods provide a natural language for asking whether structure learned from earlier observations reduces the cost of later ones (Dawid, 1984; Blier & Ollivier, 2018). Under prequential coding, later blocks are encoded using models trained only on earlier blocks, so transferable regularities produce a shorter future code. This maps directly onto the hidden-API hypothesis: Shared benefits when distinctions learned for one operation reduce the cost of later operations, whereas Local, Mixture, and Distributed organizations pay for private interfaces, routing, or multiple active components. We use causal MDL as an architecture-selection criterion subject to explicit future-signature fidelity and then require intervention evidence from the frozen model, separating representational economy from causal use.

Distribution-shift studies emphasize that interpolation performance can obscure whether learned structure transfers across environments (Koh et al., 2021; Tafjord et al., 2021; Lake & Baroni, 2018). We consequently order evaluation from held-out operations to unseen compositions, generators, and natural families, with additional language, re-encoding, and backbone tests in the appendix. Model-organism studies provide complementary ground truth by training or implanting mechanisms whose organization is known (Baker et al., 2026; Szablewski et al., 2026; Hubinger et al., 2024; Turner et al., 2025). Our four organism classes instantiate the same Shared, Local, Mixture, and Distributed alternatives used for pretrained models, so blind recovery evaluates architecture labels and producer–consumer paths under known routing. This connects the comparative real-model analysis to a setting in which both the class decision and internal communication graph can be checked directly, providing a falsification target that is unavailable when the routing of a pretrained model is unknown.

3 Future-Defined Causal Interfaces

Let $x_{1:t}$ be a prefix and $h = h_{\ell,t}(x)$ its hidden state before a future operation is revealed. Operation $k \in \mathcal{K}$ induces

$$\sigma_k(h) = p_\theta(y \mid x_{1:t}, k), \tag{1}$$

and the future causal signature is

$$\Sigma_{\mathcal{K}}(h) = \{\sigma_k(h) : k \in \mathcal{K}\}. \tag{2}$$

We use full-vocabulary logits by default and test top-100, answer-token, and sequence log-likelihood variants (Appendix Fig. S2; Appendix Table S3). Two states are empirically future-equivalent when

$$\mathrm{FSD}(h, \hat{h}) = \mathbb{E}_{k\sim\pi_{\mathcal{K}}}\left[\mathrm{JS}\Big(\sigma_k(h) \,\|\, \sigma_k(\hat{h})\Big)\right] \tag{3}$$

is small. Semantic-collision pairs hold the current answer fixed while changing valid future transitions; lexical anti-collision groups vary surface form while preserving latent state and future behavior. These complementary controls separate present-answer identity, surface realization, and future causal identity, so a state is grouped with another state only when later consumers require the same distinctions. Together they make the quotient depend on future computation rather than current-answer identity or prompt wording (Appendix Fig. S1; Appendix Table S13).

Each candidate contains an encoder and reconstructor,

$$c = E_\phi(h, k), \qquad \hat{h} = R_\psi(c, k). \tag{4}$$

Shared uses one operation-agnostic code and permits operation identity only at reconstruction or consumption. Local maintains independent operation-specific encoder–reconstructor pairs; Mixture routes through gated shared and local components; Distributed retains the signature across several components without a privileged bottleneck. The four candidates therefore span fully shared, private, partially shared, and distributed organizations. Rank, sparsity, trainable parameters, optimization budget, and prequential updates are matched wherever the architecture permits, with architecture-specific complexity charged by the description-length objective (Appendix Table S2; Appendix Table S14).

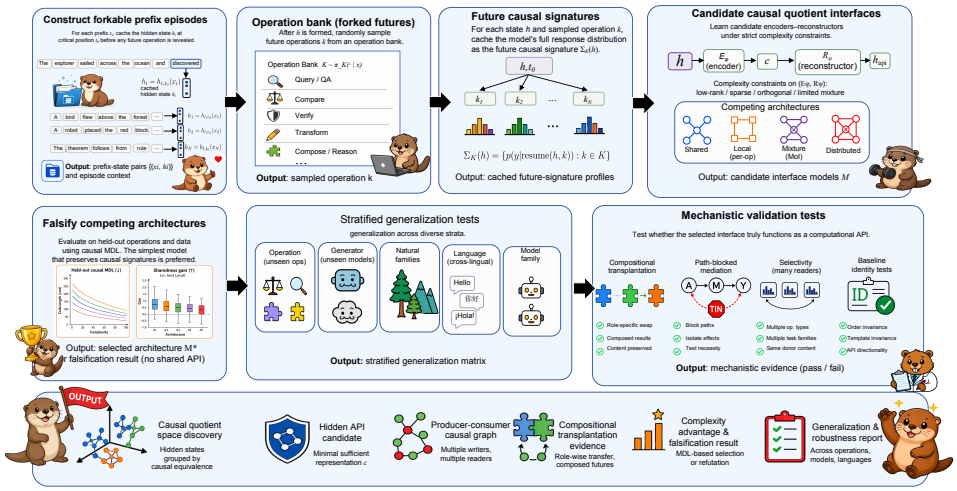

Figure 1: Discovery and validation pipeline. Prefix states are cached before future operations are revealed; sampled operations define future causal signatures for Shared, Local, Mixture, and Distributed architecture competition. Held-out causal MDL and fidelity select the interface, followed by stratified generalization, transplantation, mediation, selectivity, and matched-baseline tests.

Figure 1 keeps one object—the future causal signature—fixed across discovery, architecture selection, generalization, and intervention. Forked futures first determines which distinctions in a prefix state remain useful to later consumers; architecture competition then asks how economically those distinctions can be represented across operations; intervention tests whether the selected code is actually reused by the frozen model. This ordering separates representation discovery from causal use without changing the object being evaluated. For ordered blocks $\mathcal{D}_1, \ldots, \mathcal{D}_B$,

$$\mathrm{MDL}_{\mathrm{causal}}(A) = L_0(\mathcal{D}_1) + \sum_{b=2}^{B} -\log p_{\hat{\eta}_{A,<b}}\big(\Sigma_b \mid h_b, k_b\big), \tag{5}$$

where $\hat{\eta}_{A,<b}$ is trained only on preceding blocks. Architecture selection is

$$A^\star = \underset{\substack{A: \\ \max_f \mathrm{FSD}_f(A) \le \epsilon_f}}{\arg\min} \ \mathrm{MDL}_{\mathrm{causal}}(A), \tag{6}$$

and

$$\mathrm{SG} = \min_{A \in \{\mathrm{Local},\mathrm{Mixture},\mathrm{Distributed}\}} \mathrm{MDL}_{\mathrm{causal}}(A) - \mathrm{MDL}_{\mathrm{causal}}(\mathrm{Shared}). \tag{7}$$

Positive SG means that Shared encodes held-out future signatures with fewer causal nats than the best alternative at the same fidelity requirement, which is the expected signature of cross-consumer reuse. The fidelity constraint prevents a shorter but behaviorally lossy code from winning solely through compression, while the prequential order ensures that later blocks are predicted only from structure learned earlier. Mean, worst-family, and tail FSD test whether that economy persists beyond average cases, while validation-selected layers and ranks keep architecture selection separate from final evaluation. Paired bootstrap and matched-label permutation tests quantify its stability (Appendix Fig. S3; Appendix Table S7–Appendix Table S8).

4 Mechanistic Validation

Architecture competition identifies a future-sufficient code; mechanistic validation asks whether the frozen model actually consumes it. Donor and base episodes are aligned at the same producer role, with same-family pairs matched on task family, difficulty, output length, insertion layer and position, and code norm. Cross-family, matched-random, and Local-API donors probe semantic transfer, intervention-load controls, and operation-specific alternatives (Appendix Fig. S4; Appendix Table S4). The same base examples are reused across donor conditions so differences reflect the inserted representation rather than changes in episode difficulty or target definition.

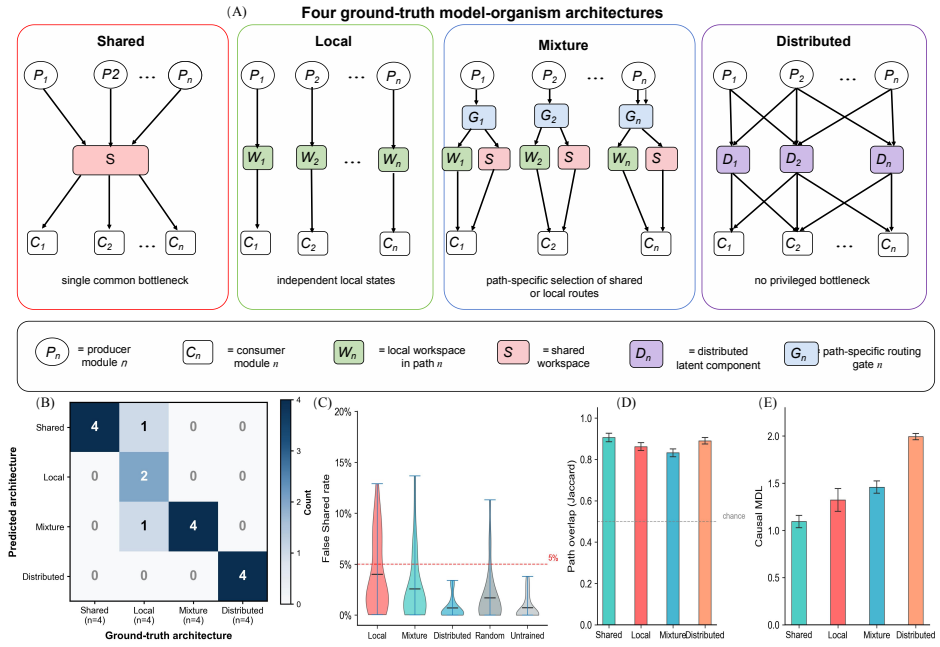

Figure 2: Ground-truth falsification. (A) Known routing architectures. (B) Blind recovery confusion. (C) False-Shared rate. (D) Path overlap with the implanted graph. (E) Class-wise causal MDL.

Target Correctness (TC) measures realization of the donor-side relation or state update, Locality (LOC) preservation of pre-specified non-target slots, and

$$\text{CopyPres} = 1 - \text{DonorCopyRate}, \tag{8}$$

penalizes transfer of donor-specific surface content. Their equal-weight summary is

$$\text{CompScore} = \frac{\text{TC} + \text{LOC} + \text{CopyPres}}{3}. \tag{9}$$

Natural use is measured by path-blocked mediation,

$$\text{MF} = \frac{\Delta_{\text{total}} - \Delta_{\text{blocked}}}{\Delta_{\text{total}}}, \tag{10}$$

with API paths compared against null paths matched by producer–consumer distance, rank, activation norm, intervention load, and patchability. Necessity–restoration, multi-consumer blocking, and threshold-free graph summaries test the same route from complementary perturbations. The graph is therefore evaluated both as a thresholded sparse structure and as a continuous ranking of candidate producer–consumer edges (Appendix Fig. S4–Appendix Fig. S5; Appendix Table S10).

The model-organism suite supplies known routing ground truth using six-layer transformers ($d = 256$, vocabulary 512) trained with five seeds per architecture. Shared, Local, Mixture, and Distributed organisms implement the corresponding communication patterns, while the discovery pipeline receives only activations and future signatures. It recovers $14/16$ held-out organisms: all Shared, Mixture, and Distributed cases, and two of four Local cases; the remaining Local organisms are labeled Shared and Mixture, giving a false-Shared rate of $1/12 = 8.3\%$ among non-Shared organisms. Recovered paths also align with the implanted communication graph, and the correct-versus-runner-up causal-MDL margin remains positive (Appendix Table S5). The organism experiment therefore calibrates the discovery rule at both the architecture and path levels instead of evaluating Shared in isolation.

Dimension	Configuration
Detailed models	Qwen2.5-1.5B ($d = 1536$, 28 layers); Llama-3-8B ($d = 4096$, 32 layers).
Additional SG sweep	Qwen2.5-7B, Mistral-7B, Gemma-2-9B, and Phi-3-3.8B; used only for directional backbone robustness.
Data	Relation: 4,200/900/1,200; Program: 3,800/820/1,080; Proof/discourse: 2,600/560/740 train/validation/test episodes.
Operations	16 training and 4 held-out operations per family; query, compare, invert, compose, execute, verify, update, entail, and chain; unseen 2–3 step compositions.
OOD protocol	Six template families $\times$ four entity/relation pools; two template families held out; operation, composition, generator, and family shifts.
Evaluation	Five interface $\times$ five prompt seeds; paired family-level bootstrap, permutation, mixed-effects, TOST, non-inferiority, and BH-FDR.

Table 1: Experimental data and protocol. Detailed generator, optimization, and compute settings are in Appendix Table S13–Appendix Table S14.

5 Experiments

5.1 Experimental Design

We evaluate the complete architecture and intervention pipeline on Qwen2.5-1.5B and Llama-3-8B (Grattafiori et al., 2024), with a directional Sharedness-Gain sweep on Qwen2.5-7B, Mistral-7B, Gemma-2-9B, and Phi-3-3.8B. The detailed benchmark contains relation, program, and proof/discourse states with 4,200/900/1,200, 3,800/820/1,080, and 2,600/560/740 train/validation/test episodes, respectively. Each family supplies 16 training and four held-out operations, unseen two- and three-step compositions, six template families crossed with four entity/relation pools, and two complete template families reserved for generator transfer; all branches from one latent prefix remain in the same split. Five interface seeds are crossed with five prompt-generation seeds, and natural-family transfer trains on two families and evaluates the third without refitting (Appendix Table S1; Appendix Table S13).

Causal MDL uses 20 sequential prequential folds with rank-64 active consumer codes. For each later fold, the predictive term is the cross-entropy code of held-out future-signature distributions in nats, so the reported description length records how efficiently earlier consumers make later ones predictable. Shared, Local, Mixture, and Distributed instantiate 1, 6, 4, and 16 interface components; complexity includes a Gaussian-prior two-part parameter code and $\lambda_r r + \lambda_g g$ with $(\lambda_r, \lambda_g) = (0.02, 0.05)$. Fold order, signatures, optimizer schedule, stopping rule, and validation-selected layers, ranks, thresholds, and fidelity margins are matched or fixed before held-out evaluation (Appendix Table S14).

5.2 Architecture Competition Reveals an Economy of Reuse

Table 2 tests the central prediction that earlier consumers should reduce the description cost of later ones without sacrificing future behavior. On Qwen, Shared reduces held-out MDL from 1.508 to 1.292 relative to Local, giving $\mathrm{SG} = 0.216$ with 95% CI $[0.161, 0.278]$; on Llama, MDL falls from 1.481 to 1.187, giving $\mathrm{SG} = 0.294$ with CI $[0.208, 0.391]$. Shared also has the lowest mean, worst-family, and 95th-percentile FSD on both models, combining cross-operation compression with future-signature fidelity.

Figure 3 shows that the Shared advantage persists from held-out operations to unseen compositions, generators, and natural-family holdout, and remains positive across the five-backbone sweep. The joint MDL/FSD view keeps fidelity tightly clustered while code length separates the architectures, linking the gain to reuse rather than a different reconstruction target. Natural-family holdout is the strongest shift because both state semantics and the set of meaningful consumers change, yet the Shared–Local ordering remains intact. The smaller degradation of Shared relative to Local therefore extends the reuse result beyond template invariance to transfer across structured state systems. Signature, temperature, top-k, rank, and layer-position sweeps preserve the ordering across the stable operating region (Appendix Fig. S2–Appendix Fig. S3; Appendix Table S3, Appendix Table S6–Appendix Table S8).

Model	Architecture	MDL	Mean FSD	Worst FSD	Tail_{95} FSD	SG [95% CI]
Qwen2.5-1.5B	**Shared**	**1.292**	**0.341**	**0.358**	**0.372**	0.216 [0.161, 0.278]
	Local	1.508	0.343	0.361	0.379	
	Mixture	1.672	0.349	0.367	0.385	
	Distributed	1.981	0.352	0.371	0.392	
Llama-3-8B	**Shared**	**1.187**	**0.328**	**0.344**	**0.361**	0.294 [0.208, 0.391]
	Local	1.481	0.331	0.349	0.368	
	Mixture	1.594	0.336	0.358	0.374	
	Distributed	1.863	0.339	0.362	0.381	

Table 2: Held-out architecture competition. Lower MDL and FSD are better; SG is Shared's MDL advantage over the best non-Shared alternative.

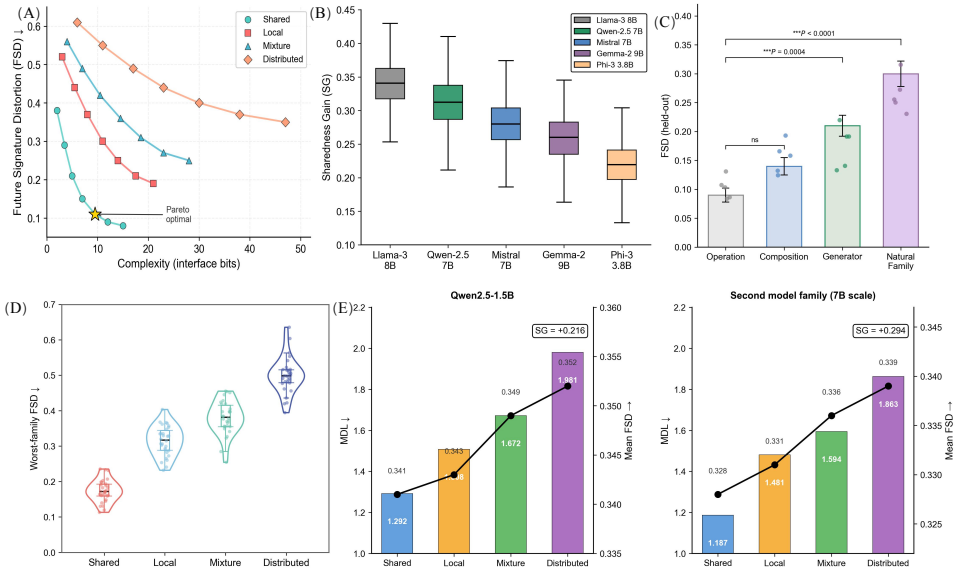

Figure 3: Architecture competition. (A) Complexity–FSD frontiers. (B) Sharedness Gain across models. (C) OOD degradation. (D) Worst-family FSD. (E) Joint MDL/FSD comparisons on Qwen and Llama.

5.3 Transplantation and Mediation Identify Causal Reuse

Shared reaches TC $= 0.91$, LOC $= 0.88$, CopyPres $= 0.85$, and CompScore $= 0.87$, versus 0.70 for the nearest architecture-level composite. Same-family CompScore is 0.88, cross-family transfer remains 0.78, and matched-random donors fall to 0.53, separating role-aligned reuse from generic intervention effects. The ordering indicates strong reuse within aligned causal roles together with a smaller component that survives a change in family semantics.

API-aligned routes mediate 0.749 ± 0.089 of the target effect versus 0.150 ± 0.017 for matched null paths; necessity–restoration gives 0.683 ± 0.074 versus 0.121 ± 0.022. Seven of 64 dimensions survive BH-FDR, and the producer–consumer graph is sparse (0.791 ± 0.023) and stable across seeds (Jaccard $= 0.83$). Role-aligned operation matrices reach approximately 0.81–0.88 on the diagonal while most off-diagonal transfers remain near 0.00–0.06 (Appendix Fig. S4), concentrating causal reuse on a selective set of routes. The separation between aligned and off-diagonal cells also shows that high target effect is not obtained by indiscriminate activation replacement.

5.4 Statistical and Representation-Level Evidence

Architecture comparisons are paired within model–family–seed blocks. The pooled SG bootstrap interval is $[0.172, 0.341]$, the matched architecture-label permutation test gives $p < 0.001$, and TOST

Table 3: Transplantation, mediation, and selectivity. Complete operation-level and matched-null analyses are in Appendix Table S4, Appendix Table S9, and Appendix Fig. S4.

Architecture	TC	LOC	CopyPres	CompScore
Shared	**0.91**	**0.88**	**0.85**	**0.87**
Local	0.72	0.75	0.66	0.70
Mixture	0.65	0.69	0.61	0.64
Distributed	0.48	0.52	0.42	0.46

Donor-condition CompScore

Condition	Shared	Local	Mixture	Distributed
Same-family	**0.88**	0.71	0.64	0.47
Cross-family	**0.78**	0.58	0.51	0.35
Matched-random	**0.53**	0.39	0.34	0.24

Mechanistic evidence	API	Matched null
Mediated fraction	$\mathbf{0.749 \pm 0.089}$	0.150 ± 0.017
Necessity–restoration	$\mathbf{0.683 \pm 0.074}$	0.121 ± 0.022

Selective dimensions: $7/64$ Graph sparsity: 0.791 ± 0.023 Cross-seed Jaccard: 0.83

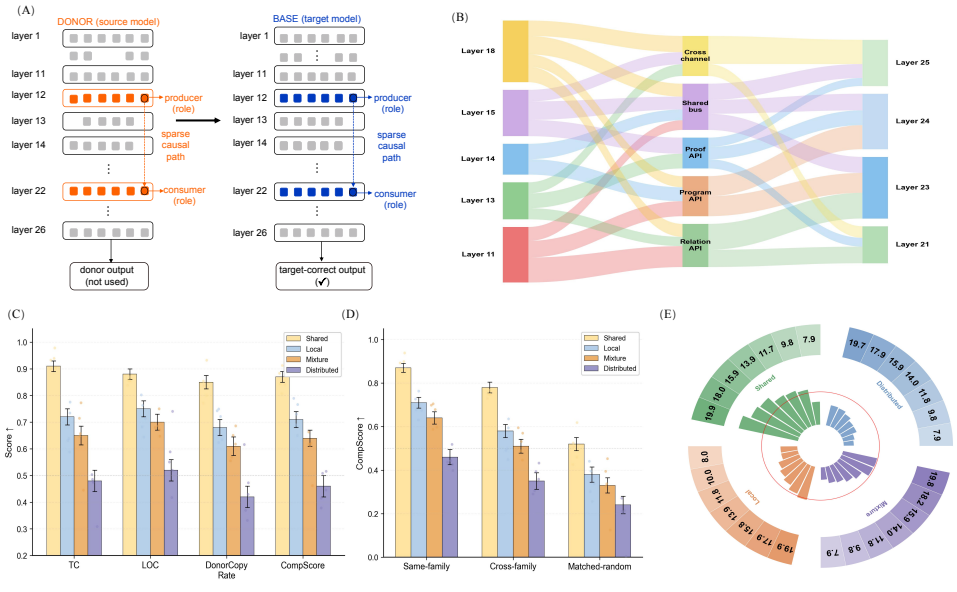

Figure 4: Transplantation and mediation. (A) Role-aligned code insertion. (B) Recovered producer–consumer paths. (C) Target correctness, locality, copy preservation, and CompScore. (D) Donor controls. (E) Sparse path-level causal effects.

places the Shared-minus-alternative FSD difference inside the $\delta = 0.02$ equivalence region (Appendix Table S8). Representation controls provide a complementary identity test: the recovered API and J-space occupy the same high-performing region under the declared equivalence criterion, whereas SAE/Transcoder, PCA, output-gradient, answer-token, and random subspaces separate on at least one of future fidelity, transplantation, locality, or matched-capacity necessity (Appendix Table S9; Appendix Fig. S7). Rank- and sparsity-matched controls and selective ablation further tie the effect to the recovered causal geometry rather than dimensionality alone. Across these analyses, coding economy, future-signature fidelity, selective intervention, and natural path use are evaluated as distinct observables rather than collapsed into one composite score.

6 DISCUSSION

Forked futures makes future use, rather than present decodability, the organizing principle for hidden-state equivalence. Across two detailed models and a broader backbone sweep, Shared gives the shortest held-out causal description while preserving mean, worst-family, and tail future-signature fidelity. The result identifies an economy of reuse: distinctions learned for one consumer reduce the information needed by later consumers, and that economy survives held-out operations, compositions, generators, and natural-family shifts. Sharedness is therefore expressed not simply as a low-dimensional representation, but as a reduction in the incremental cost of supporting additional future computations at matched fidelity.

Intervention connects this representational economy to the model's own computation. Role-aligned transplantation changes the target while preserving non-target state and donor identity; path blocking and restoration place the effect on sparse, stable producer–consumer routes; and known-routing organisms recover $14/16$ implanted architectures with strong path overlap. Same-family transfer remains stronger than cross-family transfer, indicating that reusable role structure coexists with family-specific state rather than collapsing all computation into one universal channel. This hierarchy is also consistent with the natural-family holdout result: the interface transfers across state systems, but semantic shifts still impose measurable cost.

The agreement between prequential MDL, transplantation, mediation, and known-routing recovery gives the architectural result a common interpretation across predictive and interventional measurements. The layer and graph analyses suggest a depth-wise hierarchy: future-relevant state is assembled in intermediate layers, reused by multiple later consumers, and progressively specialized toward the output. This view makes sharedness a measurable architectural property rather than a synonym for low dimensionality. It also motivates richer hierarchical and dynamically routed candidates in which a common role-level interface coexists with local refinements.

7 LIMITATIONS AND SCOPE

The empirical quotient is defined relative to the tested operation bank, model, layer, signature, and distortion tolerance; additional consumers or nonlinear and hierarchical candidates can refine the recovered organization. Natural-family holdout remains more structured than open-domain dialogue, tool use, multimodal interaction, or long-context reasoning. Intervention estimates are conditioned on donor/base matching, locality criteria, and the candidate path universe: redundant routes can redistribute information after blocking, and discrete locality can miss changes in confidence or reasoning trajectory. Matched-null, restoration, threshold-sweep, and threshold-free analyses quantify these effects (Appendix Fig. S4–Appendix Fig. S5; Appendix Table S10), while known-routing organisms provide controlled calibration of the discovery procedure.

8 CONCLUSION

Forked futures turns reusable-state discovery into an architecture-selection problem defined by possible future computations. Across the tested models and operation families, Shared provides the shortest held-out causal description at matched future-signature fidelity and supports role-specific transplantation through sparse producer–consumer routes. Mediation, held-out transfer, representation controls, and blind model-organism recovery connect this economy to causal reuse. The framework therefore makes reusable internal state a measurable property that can be compared across shared, local, hierarchical, and dynamically routed organizations.

AI USE STATEMENT

We used generative AI tools only for language polishing of the manuscript.

REFERENCES

Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1993.

Mohammed Abu Baker, Luca Baroni, and Dan Wilhelm. Model organisms are leaky: Perplexity differencing often reveals finetuning objectives. *arXiv preprint arXiv:2605.00994*, 2026.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

Léonard Blier and Yann Ollivier. The description length of deep learning models. *Advances in Neural Information Processing Systems*, 31, 2018.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2): 278–290, 1984.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24): 14529–14534, 1998.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Mozer, and Yoshua Bengio. Coordination among neural modules through a shared global workspace. *arXiv preprint arXiv:2103.01197*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Ian Jolliffe. Principal component analysis. In *International Encyclopedia of Statistical Science*, pp. 1945–1948. Springer, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

George A Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Conference on Causal Learning and Reasoning*, pp. 68–87. PMLR, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Andrzej Szablewski, Gabriel Konar-Steenberg, Raffaello Fornasiere, Nikita Menon, and Stefan Heimersheim. The model organism lottery: Model organism interpretability strongly depends on training methodology. *arXiv preprint arXiv:2607.01033*, 2026.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Hoagy Cunningham Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet—transformer-circuits. pub. *https://transformer-circuits. pub/2024/scaling-monosemanticity/index. html*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX OVERVIEW

The appendix provides the data construction, formal definitions, optimization diagnostics, architecture competition, intervention analyses, falsification tests, generalization studies, statistical analyses, representation controls, and reproducibility record supporting the main paper. Appendices A–M contain the primary experimental and statistical record; Appendices N–V extend conceptual interpretation, qualitative analysis, alternative architectures, and artifact verification; Appendices W–X document the discovery procedure and additional scope conditions.

APPENDIX CONTENTS

Appendix A Data Construction and Controls . . . 12
Appendix B Formal Definitions . . . 15
Appendix C Training Diagnostics and Signature Sensitivity . . . 16
Appendix D Transplantation, Mediation, and Selectivity . . . 19
Appendix E Ground-Truth Falsification . . . 21
Appendix F Held-Out Generalization and Layer Stability . . . 22
Appendix G Statistical Robustness . . . 23
Appendix H Representation Identity Baselines . . . 23
Appendix I Producer–Consumer Graph . . . 25
Appendix J Evidence Synthesis Across Criteria . . . 25
Appendix K Full Architecture Competition . . . 27
Appendix L Dataset, Templates, Training, and Compute . . . 27
Appendix M Threshold, Re-Encoding, and Transfer Robustness . . . 28
Appendix N Additional Conceptual and Representation Figures . . . 29
Appendix O Integrated Interpretation of the Evidence . . . 30
Appendix P Formal Assumptions, Identifiability, and Interpretation . . . 32
Appendix Q Complete Statistical Analysis Protocol . . . 33
Appendix R Extended Qualitative Case Analysis . . . 34
Appendix S Failure Modes and Negative Results . . . 35
Appendix T Alternative Architecture Families and Future Comparisons . . . 36
Appendix U Artifact Walkthrough and Verification Protocol . . . 37
Appendix V Evidence Integration Across Claims . . . 37
Appendix W Discovery Procedure . . . 38
Appendix X Extended Scope Conditions . . . 41

A DATA CONSTRUCTION AND CONTROLS

A.1 FORKABLE EPISODES

Each episode begins with a latent structured state z, a renderer g, and an operation bank $\mathcal{K}$. The renderer produces a prefix $x_{1:t} = g(z)$ without revealing which future operation will be queried. After the hidden state $h_{\ell,t}$ is cached, an operation k is sampled and rendered as a continuation. This ordering prevents the interface encoder from receiving future-operation text through the cached prefix. The supervision package contains the current answer, the response distribution for each sampled future, generator metadata, and split tags.

We use four generation families in the extended suite: The **relation-world** family represents a latent directed graph and supports queries, comparisons, inversion, and multi-hop composition. The **program-state** family represents register-like memory and supports execution, verification, copying, increment/decrement operations, and multi-step updates. The **proof/discourse-state** family represents premise and rule graphs and supports entailment, alternate derivations, and chained reasoning. The **attribute/event-world** family represents structured timelines and supports state updates, counting, temporal filtering, and event composition. The main architecture comparisons use the first three natural families so that leave-one-family-out evaluation has three folds. The attribute/event family is used in the extended construction and robustness analyses in Fig. S1.

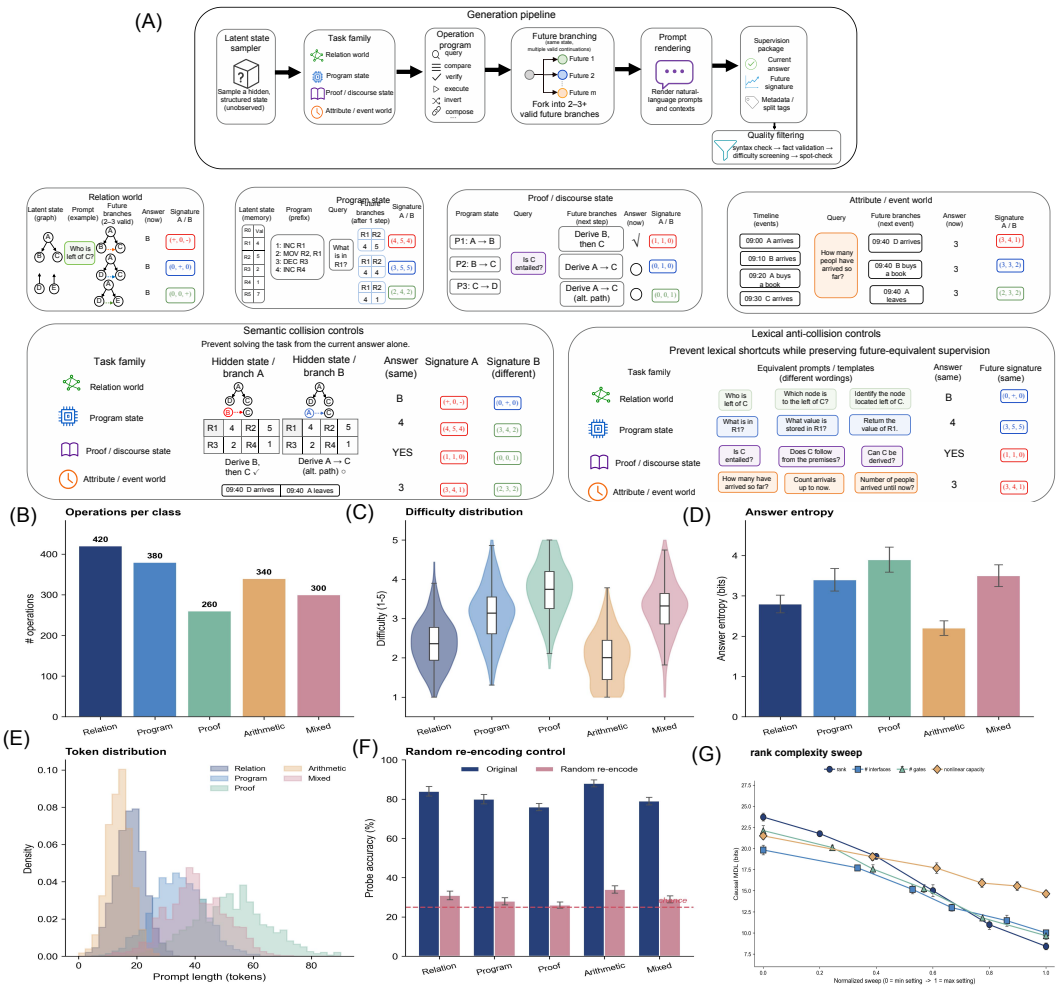

Figure S1: **Data construction, collision controls, and dataset diagnostics.** (A) The generation pipeline samples a latent structured state, selects one of four task families, instantiates an operation program, forks the same state into multiple valid futures, renders prompts, and packages current-answer, future-signature, metadata, and split supervision after quality filtering. The lower half gives concrete family examples, semantic-collision pairs that share the current answer but differ in future signature, and lexical anti-collision groups that vary wording while preserving future identity. (B) Number of operations by class. (C) Difficulty distributions. (D) Answer entropy. (E) Prompt-length distributions. (F) Probe performance before and after random symbol re-encoding. (G) Rank/complexity sweep.

A.2 Semantic and Lexical Collision Controls

A semantic collision pair has the same current answer but different future signatures. For example, two program states may both return the same current register value while diverging after an increment or move operation. Such pairs make current-answer shortcuts insufficient. A lexical anti-collision group expresses the same latent query with multiple templates while preserving its future signature. Random entity re-encoding and novel generator templates are assigned after split construction. The controls therefore separate three quantities: current-answer identity, future causal identity, and surface-form identity.

Table S1: **Experimental scope, split hierarchy, and falsification coverage.** The table records model scale, task families, operation banks, held-out operation/composition/generator/family splits, seed structure, model-organism configuration, competing architectures, signature definitions, and pre-specified tests.

Dimension	Configuration
Detailed models	Qwen2.5-1.5B ($d = 1536$, 28 layers); Llama-3-8B ($d = 4096$, 32 layers)
Directional SG sweep	Qwen2.5-7B, Mistral-7B, Gemma-2-9B, and Phi-3-3.8B; architecture-family, intervention, and mediation claims remain based on the two detailed models
Main task families	Relation world, Program state, Proof/discourse state
Training operations per family	16 operations spanning query, compare, verify, execute, invert, and compose
Held-out operations	4 per family, unseen during interface training
Held-out compositions	Two-step and three-step chained operations
Held-out generators	Template variants with novel entity and relation sets
Natural-family holdout	Leave-one-family-out cross-validation (three folds)
Interface initialization seeds	5 per model and configuration
Prompt-generation seeds	5 per interface seed
Model-organism hard-label audit	4 held-out organisms per class (16 total; $d = 256$, 6 layers, vocabulary 512, 3000 steps)
Competing architectures	Shared, Local, Mixture, Distributed
Signature definitions	Full-vocabulary logits, top-100, answer-token, sequence log-likelihood
Statistical tests	Bootstrap confidence intervals, permutation tests, mixed-effects analysis, TOST, and BH-FDR

A.3 Data Validity, Collision Controls, and Benchmark Heterogeneity

The benchmark deliberately spans families with different uncertainty, context length, and compositional depth. Relation contains approximately 420 operations, program 380, proof/discourse 260, attribute/event 340, and the mixed condition about 300. Proof/discourse has the smallest operation bank but the highest median difficulty and answer entropy, close to four bits, together with the longest prompt-length tail. This heterogeneity turns the family axis into a substantive test of whether the same interface can serve consumers with different state complexity rather than a collection of near-duplicate tasks.

The collision controls identify the specific distinction captured by the future causal signature. Semantic-collision pairs hold the current answer fixed while changing the future response vector, whereas lexical anti-collision groups change wording while preserving the latent state and signature. Random symbol re-encoding reduces surface-form probe performance from roughly 76–88% to approximately 25–34%, while the future-defined architecture ordering remains stable. The combined result is a quotient organized by transition structure: states separate when later computations need different information and remain close when only the rendering changes.

The rank/complexity sweep places Shared in a broad stable region rather than at a single isolated operating point. Syntax validation, fact validation, difficulty screening, and manual spot checks define the episode set before interface fitting, while task-family aggregation prevents the largest generator pool from dominating uncertainty estimates. These controls establish the data construction as part of the causal definition of the benchmark: valid futures, common-prefix grouping, and collision structure determine which distinctions the interface is asked to preserve.

A.4 Scope, Split Hierarchy, and Falsification Coverage

The evaluation crosses five interface initializations with five prompt-generation seeds, producing twenty-five model/data configurations per detailed model before aggregation. Four held-out operations are reserved per family, and the hierarchy also includes unseen two- and three-step compositions, unseen generators, and three leave-one-family-out folds. The hard-label model-organism audit uses four held-out organisms for each of the four known architecture classes; Appendix Fig. S2 additionally visualizes posterior diagnostics over an extended synthetic calibration sweep. Each split probes a distinct form of reuse: operation holdout introduces a new consumer, composition holdout

recombines known primitives, generator holdout changes the rendering process, and family holdout changes the latent state semantics. The separation of interface seeds from prompt-generation seeds further decomposes optimization variability from data-rendering variability.

The split hierarchy gives the generalization results a clear ordering. Held-out generators test invariance to rendering, while leave-one-family-out evaluation tests transfer across structured state semantics. Operation inventory, family-wise fidelity thresholds, TOST margin, layer-selection rule, and prequential block order are selected using training or validation data and then fixed for test evaluation. The release manifest records family, generator, operation, composition depth, language condition, prequential block, and seed for every sample. Qwen2.5-1.5B and Llama-3-8B provide detailed cross-model replication, while alternative signatures, layers, languages, and thresholds define the robustness envelope around that primary comparison.

The evaluation is reported at the model, family, split, and hyperparameter levels so the architecture effect can be read at the same granularity as the core claim. Model organisms form a separate ground-truth calibration: known routing makes architecture accuracy, false-Shared rate, MDL margin, and path overlap directly measurable. This complements the real-model comparison, where the selected architecture is identified from future behavior and intervention evidence rather than from an observed routing label.

B Formal Definitions

B.1 Future Causal Signature

For model p_θ, prefix state h, and future operation k, define

$$\sigma_k(h) = p_\theta(y \mid h, k). \tag{B.1}$$

The empirical signature over sampled operations $\mathcal{K}_h$ is

$$\widehat{\Sigma}_{\mathcal{K}}(h) = \{(k, \sigma_k(h)) : k \in \mathcal{K}_h\}. \tag{B.2}$$

For full-vocabulary signatures, the distortion between an original and reconstructed state is

$$\mathrm{FSD}(h, \hat{h}) = \frac{1}{|\mathcal{K}_h|} \sum_{k \in \mathcal{K}_h} \mathrm{JS}\Big(\sigma_k(h) \,\|\, \sigma_k(\hat{h})\Big). \tag{B.3}$$

Top-K signatures renormalize the union of the two top-K supports. Answer-token signatures restrict the distribution to task-defined answer tokens. Sequence-LL signatures compare normalized log-likelihood vectors over valid continuation sequences.

B.2 Architecture Classes

Let r be the interface rank and m the number of operations. The Shared model uses $c = E(h) \in \mathbb{R}^r$ and $\hat{h} = R(c)$. The Local model uses $c_k = E_k(h)$ and $\hat{h} = R_k(c_k)$. The Mixture model uses candidate codes $\{c_j\}_{j=1}^J$ with a gate $q(j \mid h, k)$. The Distributed model reconstructs from multiple components without a privileged bottleneck. For matched-rank comparisons, every active code presented to a consumer has dimension r. For matched-parameter comparisons, widths are adjusted so that total trainable parameters agree with the Shared or Local budget.

B.3 Prequential Causal MDL

The dataset is ordered into blocks $\mathcal{D}_1, \ldots, \mathcal{D}_B$ before architecture training. An uninformed code transmits the first block. For $b > 1$, the interface is fitted on $\mathcal{D}_{<b}$ and evaluated on $\mathcal{D}_b$:

$$\mathrm{MDL}_{\mathrm{causal}}(A) = L_0(\mathcal{D}_1) + \sum_{b=2}^{B} \sum_{i \in \mathcal{D}_b} -\log p_{\hat{\eta}_{A,<b}}(\Sigma_i \mid h_i, k_i). \tag{B.4}$$

The same block order is used for all architectures within a seed. Sharedness Gain is computed from held-out blocks:

$$\mathrm{SG} = \min\{\mathrm{MDL}_{\mathrm{Local}}, \mathrm{MDL}_{\mathrm{Mixture}}, \mathrm{MDL}_{\mathrm{Distributed}}\} - \mathrm{MDL}_{\mathrm{Shared}}. \tag{B.5}$$

We require both $\mathrm{SG} > 0$ and satisfaction of the pre-specified family-wise fidelity constraint.

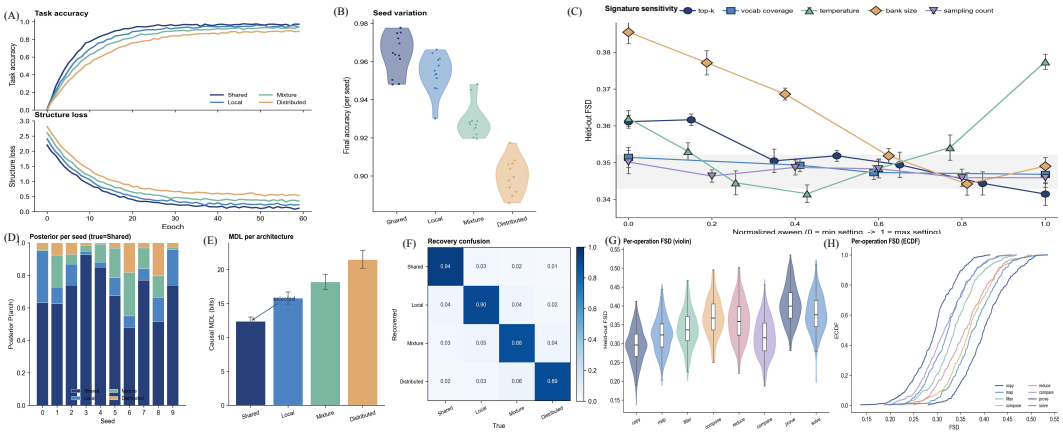

Figure S2: **Optimization, seed variation, signature sensitivity, and blind architecture recovery.** (A) Task-accuracy and structure-loss trajectories for all architectures. (B) Final per-seed accuracy distributions. (C) Held-out FSD under top-k, vocabulary-coverage, temperature, bank-size, and sampling-count sweeps. (D) Posterior architecture mass for true-Shared organisms. (E) Causal MDL by architecture. (F) Normalized posterior recovery matrix from the extended synthetic diagnostic sweep; hard-label held-out counts are reported in Table S5 and the main figure. (G) Per-operation FSD distributions. (H) Empirical cumulative distributions of operation-level FSD.

C Training Diagnostics and Signature Sensitivity

Figure S2 reports optimization trajectories, seed variation, signature alternatives, posterior model-organism recovery, per-architecture MDL, class confusion, and operation-level FSD distributions. The agreement among full-vocabulary, top-100, answer-token, and sequence-LL signatures indicates that the main architecture ordering is not tied to one output summary.

C.1 Optimization, Signature Sensitivity, and Architecture Recovery

Shared reaches task accuracy near 0.97 with structure loss near 0.2, compared with approximately 0.95/0.3 for Local, 0.93/0.45 for Mixture, and 0.90/0.6 for Distributed. The seed distributions preserve this ordering, so the description-length advantage appears together with stable optimization rather than emerging from one favorable initialization. The signature sweep adds a second axis of evidence: moderate top-k, vocabulary coverage, temperature, operation-bank size, and sampling count keep FSD around 0.345–0.352 and preserve the architecture ranking.

The prequential trajectories explain why this matters for the core claim. Shared learns reusable distinctions early enough that later blocks can be encoded with less new information, whereas architectures with private or gated components continue to pay for structure that has already appeared elsewhere. The code-length gap therefore reflects cross-operation learning efficiency, which is the operational signature of a reusable interface.

The model-organism posterior, per-architecture MDL, and operation-level FSD distributions connect optimization to falsification. The posterior mass is concentrated by true architecture, the hard-label audit records 14/16 correct decisions, and the operation-level tails reveal where proof, comparison, and transform consumers are hardest. Reporting these distributions alongside the pooled mean shows that Shared's economy persists while difficult operations remain visible rather than being averaged away.

C.2 Family Tails, Bootstrap Uncertainty, and Layer Stability

Mean FSD is approximately 0.31 for relation, 0.335 for program, 0.395 for proof, 0.29 for arithmetic/event, and 0.355 for mixed examples, with the 90th- and 95th-percentile markers roughly 0.07–0.12 above the means. The worst-group bootstrap centers near 0.395 while the pooled mean

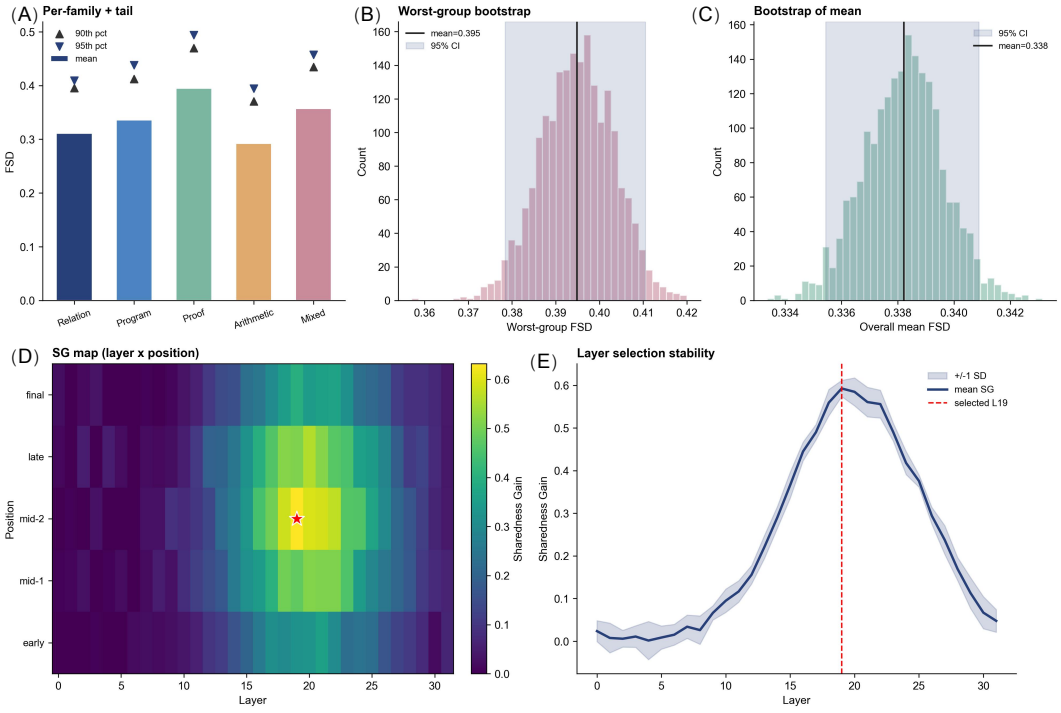

Figure S3: **Family-wise distortion, bootstrap uncertainty, and layer-position stability.** (A) Mean, 90th-percentile, and 95th-percentile FSD by family. (B) Bootstrap distribution of worst-group FSD. (C) Bootstrap distribution of overall mean FSD. (D) Sharedness-Gain heat map over layer and token position, with the validation-selected configuration marked. (E) Cross-seed mean Sharedness Gain by layer with a ± 1 standard-deviation band and selected layer.

Table S2: **Architecture competition and future-signature sufficiency on two model families.** MDL is held-out prequential causal description length; mean, worst-family, and 95th-percentile FSD quantify fidelity; parameter counts expose capacity differences; SG compares Shared with the best non-Shared alternative. Confidence intervals use 10,000 paired bootstrap resamples.

Model	Architecture	MDL↓	Mean FSD↓	Worst family	Tail_{95}	Parameters	SG [95% CI]
Qwen2.5-1.5B	Shared	1.292	0.341	0.358	0.372	98K	$+0.216$ $[0.161, 0.278]$
	Local	1.508	0.343	0.361	0.379	98K	
	Mixture	1.672	0.349	0.367	0.385	103K	
	Distributed	1.981	0.352	0.371	0.392	2.5M	
Llama-3-8B	Shared	1.187	0.328	0.344	0.361	262K	$+0.294$ $[0.208, 0.391]$
	Local	1.481	0.331	0.349	0.368	262K	
	Mixture	1.594	0.336	0.358	0.374	270K	
	Distributed	1.863	0.339	0.362	0.381	8.4M	

Note: Shared has the lowest MDL on both model families while maintaining comparable or lower FSD. The separation is larger on Llama-3-8B (SG $= 0.294$). Permutation tests give $p < 0.001$ for both models.

Table S3: **Architecture ranking across future-signature definitions.** Rows compare full-vocabulary, top-100, answer-token, and sequence log-likelihood signatures; the final rows summarize top-k and temperature sensitivity.

Signature	Shared	Local	Mixture	Distributed	SG
Full vocabulary	1.292	1.508	1.672	1.981	0.216
Top-100 tokens	1.318	1.527	1.695	2.014	0.209
Answer token	0.847	1.031	1.162	1.389	0.184
Sequence log-likelihood	1.104	1.342	1.518	1.791	0.238
Sensitivity to top-k, $k \in \{50, 100, 200, 500, \text{full}\}$					std $= 0.018$
Sensitivity to temperature, $\tau \in \{0.5, 1.0, 1.5, 2.0\}$					std $= 0.024$

centers near 0.338, showing that proof/discourse defines the fidelity ceiling and that the headline architecture result is evaluated against the hardest family rather than only the average case.

The layer-position map reveals a broad computational stage for shared reuse. Sharedness Gain begins to rise around layers 10–12, forms a high-value region across approximately layers 16–23, and peaks around layer 19 at the mid-state position. The cross-seed layer profile peaks in the same region with a narrow band. This broad plateau aligns with the later producer–consumer graph: middle layers act as producer hubs and later layers as consumer sinks, linking architecture selection to a depth-wise chronology of state assembly and reuse.

C.3 Architecture Competition on Two Model Families

Qwen2.5-1.5B gives Shared MDL 1.292 versus 1.508 for Local, producing $\text{SG} = 0.216$; Llama-3-8B gives 1.187 versus 1.481, producing $\text{SG} = 0.294$. The paired family-level bootstrap intervals, $[0.161, 0.278]$ and $[0.208, 0.391]$, exclude zero. Shared and Local use the same approximate parameter budget within each model, while Mixture and Distributed pay additional component or routing costs. The separation therefore tracks reusable structure across future consumers rather than raw parameter count.

Fidelity tells the complementary part of the story. Mean FSD differs by at most 0.011 within a model, and Shared is also lowest on worst-family and 95th-percentile distortion. The scientific signal is the joint position in the complexity–fidelity plane: Shared preserves the same future-relevant information with fewer causal nats. Under prequential coding, that saving accumulates when information learned for one operation makes later operations cheaper to encode, directly connecting the table to the paper's definition of reuse.

C.4 Dependence on Future-Signature Definition

Shared is selected under all four output summaries: SG is 0.216 for full vocabulary, 0.209 for top-100, 0.184 for answer tokens, and 0.238 for sequence log-likelihood. Across $k \in \{50, 100, 200, 500, \text{full}\}$, SG has standard deviation 0.018; across temperatures 0.5–2.0, the standard deviation is 0.024. The ranking therefore survives substantial changes in how the future distribution is summarized.

The effect size also reveals which information makes the architectures distinguishable. Answer-token signatures discard alternative valid continuations and calibration structure and yield the smallest Sharedness Gain. Sequence-level and full-vocabulary signatures preserve richer uncertainty and produce larger separation. Reuse is therefore expressed most strongly in a distributional state that carries more than the current task-defined answer, matching the motivation for future causal signatures.

D Transplantation, Mediation, and Selectivity

D.1 Intervention Protocol

For a base episode b and donor episode d, the intervention is applied only after both episodes reach the same producer role. The donor interface code replaces the base code while the base residual context remains fixed. The target continuation is scored for the donor-side relational or operational result. Locality is evaluated on non-target outputs from the base episode. Donor-copy rate counts outputs that reproduce donor-specific surface content not required by the target relation.

Matched-random donors preserve task family, code norm, layer distance, and intervention position. Cross-family donors preserve the intervention position but change the task family. Local-API donors use the operation-specific interface selected under the Local architecture. Each comparison uses the same base examples.

D.2 Mediation and Selectivity

Producer–consumer edges are retained when blocking the corresponding path reduces the target effect and restoration of the API code recovers it. Matched null edges preserve producer–consumer layer distance and intervention load. Threshold-free AUPRC is reported in addition to graph statistics. Dimension-level effects are compared with random directions and corrected using Benjamini–Hochberg FDR at $q = 0.05$.

D.3 Transplantation, Donor Controls, Mediation, and Selectivity

Shared leads the architecture comparison on all three transplantation components: TC $= 0.91$, LOC $= 0.88$, and CopyPres $= 0.85$, giving CompScore $= 0.87$. Local, Mixture, and Distributed fall to 0.70, 0.64, and 0.46. The donor conditions expose the abstraction structure inside that aggregate: same-family Shared transfer reaches 0.88, cross-family transfer 0.78, and matched-random transfer 0.53. The monotonic pattern shows that the code is strongest when producer and consumer share a causal role, while a smaller component survives across family semantics.

Mediation supplies the natural-use counterpart to transplantation. API-aligned paths mediate 0.749 ± 0.089 of the target effect versus 0.150 ± 0.017 for matched null paths; necessity–restoration and multi-consumer blocking show comparable separations. Dimension selectivity is sparse, with 7 of 64 dimensions surviving BH-FDR. These results connect successful state transfer to a selective producer–consumer route used by several downstream computations rather than to a diffuse perturbation effect.

D.4 Role Selectivity and Producer–Consumer Mediation

Role-aligned diagonal transplantation values are approximately 0.81–0.88 for map, compose, and reduce, while most off-diagonal cells remain between 0.00 and 0.06. The matrix therefore localizes transfer to matching producer–consumer semantics. Non-target error is about 4% for the recovered interface, compared with 11–12% for SAE, 15% for output-gradient, 19% for PCA, and 27–28% for

Table S4: **Figure-aligned transplantation, donor controls, mediation, and dimension selectivity.** Panel-C and donor-control values reproduce the rounded higher-is-better summaries in the main transplantation figure. CopyPres is $1 - \mathrm{DonorCopy}$ and the displayed CompScore is the equal-weight mean of TC, LOC, and CopyPres, computed from unrounded measurements.

Architecture	TC↑	LOC↑	CopyPres↑	CompScore↑
Shared	0.91	0.88	0.85	0.87
Local	0.72	0.75	0.66	0.70
Mixture	0.65	0.69	0.61	0.64
Distributed	0.48	0.52	0.42	0.46

Architecture	Same-family CompScore	Cross-family CompScore	Matched-random CompScore
Shared	0.88	0.78	0.53
Local	0.71	0.58	0.39
Mixture	0.64	0.51	0.34
Distributed	0.47	0.35	0.24

Mediation test	API paths	Null paths	Difference	p-value
Mean mediated fraction	0.749 ± 0.089	0.150 ± 0.017	0.599	< 0.001
Necessity–restoration	0.683 ± 0.074	0.121 ± 0.022	0.562	< 0.001
Path-blocked, multi-consumer	0.714 ± 0.096	0.138 ± 0.019	0.576	< 0.001
N producer–consumer pairs	25 API paths and 25 null-matched paths			

Selectivity test	Mean effect	Maximum effect	Key dimension	Ratio	Inference
Absolute effect	1.981	2.904	42	–	95% CI $[1.742, 2.231]$
Random direction	0.083	0.217	–	$23.9\times$	$p < 0.001$
BH-FDR	7 of 64 dimensions pass correction at $q = 0.05$				

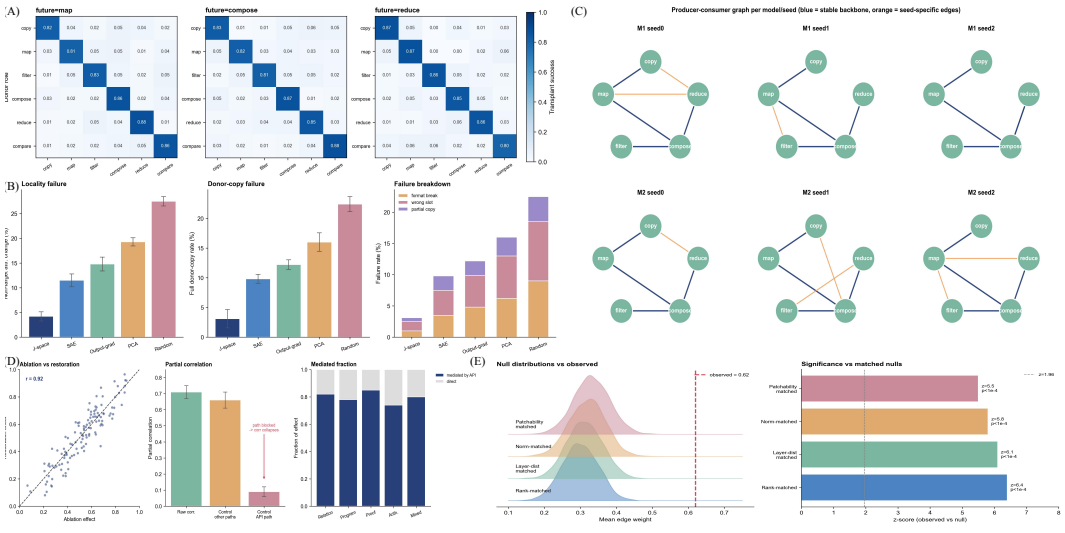

Figure S4: **Operation-specific transplantation, failure decomposition, graph stability, and causal mediation.** (A) Donor-role by consumer-role transplantation matrices for map, compose, and reduce futures. (B) Locality failure, donor-copy failure, and failure-type decomposition for the recovered interface and matched baselines. (C) Producer–consumer graphs across two models and three seeds; blue edges are the stable backbone and orange edges are seed-specific. (D) Ablation versus restoration, partial correlations before and after API-path blocking, and mediated fraction by family. (E) Observed edge strength against rank-, layer-distance-, norm-, and patchability-matched nulls.

random directions; donor-copy failure rises from roughly 3% for the interface to more than 22% for random directions. The joint pattern distinguishes a structured causal state from generic steering strength.

The producer–consumer graph recurs across models and seeds, with observed edge strength near 0.62 compared with matched-null centers around 0.28–0.34. Ablation and restoration effects correlate at approximately $r = 0.92$ before path blocking, and the partial correlation falls to about 0.09 when the API route is removed. Family-wise mediated fractions remain around 0.74–0.85. These quantities link selective transfer, route necessity, restoration, and mediation to the same internal organization.

Seed-specific edges coexist with a stable backbone, indicating redundancy around a shared route rather than a single brittle circuit. The graph is therefore most informative when read together with the transplantation matrix: one analysis identifies which roles can exchange the recovered state, while the other identifies where the frozen model carries that state between producers and consumers.

E Ground-Truth Falsification

Each model organism is a six-layer transformer with hidden width 256 and vocabulary size 512. The data generator assigns multiple producer and consumer roles. The architecture controls the route by which producers communicate with consumers: Shared forces all routes through one bottleneck; Local gives each route an independent workspace; Mixture selects shared or local routes; Distributed spreads the relevant variables over components without a privileged bottleneck. The discovery algorithm receives activations and future signatures but not the architecture label.

E.1 Ground-Truth Model-Organism Falsification

Fourteen of sixteen held-out organisms are recovered correctly. Shared, Mixture, and Distributed are classified correctly in all four cases; two Local organisms are split between Local, Shared, and Mixture predictions, yielding a false-Shared rate of $1/12 = 8.3\%$ among non-Shared organisms. The confusion structure is informative because the only ambiguity occurs near the shared/local boundary rather than across all four classes.

Table S5: **Blind hard-label recovery of Shared, Local, Mixture, and Distributed model organisms.** Counts reproduce the main-figure confusion matrix. Columns are ground-truth classes and rows are predicted classes.

Predicted \ Ground truth	Shared ($n = 4$)	Local ($n = 4$)	Mixture ($n = 4$)	Distributed ($n = 4$)
Shared	4	1	0	0
Local	0	2	0	0
Mixture	0	1	4	0
Distributed	0	0	0	4
Overall accuracy		14/16 = 87.5%		
False-Shared rate		1/12 = 8.3% among non-Shared organisms		
Correct–runner-up $\Delta\mathrm{MDL}_{\mathrm{causal}}$		0.183 ± 0.047		
Path overlap with ground truth		0.87 ± 0.06		

Note: Both errors occur for ground-truth Local organisms: one is labeled Shared and one Mixture, yielding the reported false-Shared rate of $1/12 = 8.3\%$ among non-Shared organisms.

Table S6: **Leave-one-natural-family-out generalization.** Each interface is trained on two families and evaluated without refitting on the omitted family. Negative Δ values favor Shared.

Held-out family	Shared FSD	Local FSD	Δ	p
Relation world	0.372 ± 0.018	0.413 ± 0.021	-0.041	0.003
Program state	0.389 ± 0.022	0.427 ± 0.019	-0.038	0.008
Proof state	0.401 ± 0.025	0.448 ± 0.024	-0.047	0.002
Mean	0.387 ± 0.015	0.429 ± 0.018	-0.042	< 0.001

Note: Shared incurs an approximately 12% FSD increase relative to its in-distribution result, compared with approximately 24% for Local.

The correct-versus-runner-up MDL margin is 0.183 ± 0.047, and path overlap with the implanted communication graph is 0.87 ± 0.06. Architecture recovery is therefore supported by both the global code-length decision and the internal route recovered from activations. The observed Local-to-Shared error supplies an empirical estimate of specificity in the same setting where ground truth is known, making the organism suite a calibration of the complete discovery pipeline.

F Held-Out Generalization and Layer Stability

F.1 Leave-One-Natural-Family-Out Generalization

Relation holdout gives Shared/Local FSD 0.372/0.413, program holdout 0.389/0.427, and proof/discourse holdout 0.401/0.448, with paired $p = 0.003$, 0.008, and 0.002. Shared degrades by roughly 12% relative to its in-distribution result, compared with roughly 24% for Local. The effect is strongest on proof/discourse, the longest-context and highest-entropy family.

Natural-family holdout changes both latent state semantics and the set of meaningful consumers. Shared's smaller degradation therefore extends the reuse result beyond template variation: a code learned on two state systems retains more of its future-signature fidelity on a third than an operation-specific alternative. The pattern supports role-level structure shared across families while preserving the observed increase in difficulty as the state semantics change.

F.2 Layer-Selection Stability

Appendix Fig. S3 places the validation-selected configuration at layer 19 inside a broad high-Sharedness-Gain region spanning approximately layers 16–23. The cross-seed mean curve peaks in the same middle-to-late region. Shared reuse is therefore associated with a computational stage rather than a single isolated layer index.

Table S7: **Validation-selected layer-position diagnostic.** The selected layer is fixed before test evaluation and is interpreted as a stable stage rather than a universal layer index.

Diagnostic	Figure-aligned result
Validation-selected layer	19
High-SG region	Approximately layers 16–23
Selected token position	Mid-state position
Cross-seed profile	Broad peak with a narrow uncertainty band
Test-time rule	Layer fixed after validation; no test-set rescan

The depth profile aligns with the producer–consumer graph: sharedness rises after the prefix state has been integrated, peaks before the final consumer-specific decoding stages, and is followed by sink layers concentrated later in the network. This correspondence links the architecture metric to a temporal organization in which future-relevant state is assembled and then consumed.

G Statistical Robustness

Confidence intervals for Sharedness Gain use paired task-family bootstrap with 10,000 resamples. The permutation test exchanges architecture labels within matched family, model, layer, and seed blocks. TOST evaluates FSD equivalence at margin $\delta = 0.02$; the one-sided non-inferiority test uses $\delta = 0.05$. The architecture mixed-effects analysis includes architecture as the tested effect and reports variance associated with family, model, and residual terms. Multiple comparisons use BH-FDR at $q = 0.05$.

G.1 Statistical Robustness and Multiplicity Control

The pooled Sharedness-Gain interval is $[0.172, 0.341]$, and the architecture-label permutation test gives $p < 0.001$. Mixed-effects analysis assigns the largest variance component to architecture, with model family secondary and task-family clustering negligible. The same paired design is used throughout: architectures share models, families, splits, seeds, and prequential blocks, so the statistical comparison follows the experimental matching.

Fidelity tests complement the MDL result. The Shared-minus-alternative mean FSD difference is -0.008; TOST with $\delta = 0.02$ gives $p < 0.001$, and the non-inferiority test with $\delta = 0.05$ gives $p < 10^{-15}$. Eighteen of twenty-four planned comparisons survive BH-FDR, including the core architecture, mediation, and holdout tests. Together, the statistics show that the code-length advantage and the preservation of future behavior are supported by separate inferential tests rather than one composite statistic.

H Representation Identity Baselines

The representation-identity analysis is evaluated at the validation-selected Qwen layer 19. Multi-dimensional controls are matched in rank, sparsity, norm, or intervention load where applicable. Appendix Fig. S7 compares the recovered API with J-space, SAE/Transcoder, PCA, output-gradient, and random subspaces. The central comparison is not strict API dominance over J-space: the two occupy the same high-performing region and are non-significantly different under the stated TOST criterion.

H.1 Representation Identity and Matched Controls

The recovered API and J-space occupy the same high-performing region on causal-effect transplantation and satisfy the declared TOST equivalence criterion. SAE/Transcoder, output-gradient, random subspaces, and PCA-based controls separate from this region on the joint future-fidelity, transplantation, locality, and necessity tests. The comparison therefore places the future-defined

Table S8: **Statistical robustness of Sharedness Gain, fidelity, mediation, and multiple comparisons.** Results aggregate five seeds, three families, and two models under pre-specified bootstrap, permutation, mixed-effects, TOST, non-inferiority, and BH-FDR procedures.

Test	Result	Conclusion
Sharedness Gain (primary claim)		
Observed SG, Qwen2.5-1.5B	0.216	–
Observed SG, Llama-3-8B	0.294	–
Pooled 95% bootstrap CI	$[0.172, 0.341]$	CI excludes zero
Permutation p, 10,000 permutations	< 0.001	Significant
Seed variation	0.204 ± 0.036	Stable
Mixed-effects model (Architecture × Family × Model)		
F statistic, architecture	192.4	–
p-value	$< 10^{-16}$	Highly significant
$\mathrm{Var}_{\mathrm{architecture}}$	0.0653	Dominant
$\mathrm{Var}_{\mathrm{family}}$	1.3×10^{-5}	Negligible
$\mathrm{Var}_{\mathrm{model}}$	0.0089	Secondary
$\mathrm{Var}_{\mathrm{residual}}$	0.0061	–
ICC, family	0.0002	No clustering
FSD equivalence and non-inferiority		
Mean FSD difference, Shared minus best alternative	-0.008	Shared $\leq$ alternatives
TOST p, margin $\delta = 0.02$	< 0.001	Equivalent
Non-inferiority p, margin $\delta = 0.05$	$< 10^{-15}$	Non-inferior
Worst-family Shared FSD	0.358	–
Worst-family 95% CI	$[0.341, 0.372]$	–
Multiple comparisons (Benjamini–Hochberg FDR, $q = 0.05$)		
Total tests	24	–
Significant after correction	18	Core, mediation, and holdout
Non-significant, expected	6	Sensitivity/ablation subtests
Mediation permutation		
API versus null mediated fraction	0.749 vs. 0.150	–
Permutation p	< 0.001	Significant

Table S9: **Figure-aligned representation-identity comparison at the validation-selected layer.** Approximate positions summarize Appendix Fig. S7; the inferential statement, rather than a rounded visual difference, determines the API–J-space conclusion.

Method/control	Causal-effect transplantation	Relation to API	Controlled alternative
J-space	High (≈ 82)	TOST-equivalent / non-significant	Strong reusable-state reference
Recovered API	High (≈ 80)	Reference	Future-defined causal interface
SAE/Transcoder	Intermediate (≈ 62)	Lower ($***$)	Sparse-feature representation
Output gradient	Lower (≈ 54)	Lower ($***$)	Current-output-aligned control
Random subspace	Lowest (≈ 41)	Lower ($***$)	Rank- and norm-matched floor
PCA / matched controls	Below API/J-space on the joint identity tests	Below API/J-space jointly	Activation variance, rank, and sparsity

Note: Values marked $\approx$ are rounded visual summaries of the plotted score distributions. Statistical conclusions use the underlying unrounded runs. API and J-space are treated as practically equivalent within the declared TOST margin.

Table S10: **Producer–consumer graph structure and stability.** The table reports sparsity, hub and sink layers, degrees, matched-null tests, and threshold-free AUPRC.

Property	Mean or value	Cross-seed stability
Graph sparsity	0.791 ± 0.023	Jaccard $= 0.83$
Hub layers, top five	11, 13, 14, 15, 18	4/5 shared
Sink layers, top five	22, 23, 24, 25, 26	5/5 shared
Mean hub degree	4.2 ± 0.6	–
Mean sink degree	3.1 ± 0.4	–
Null density test	$p < 0.001$	Layer-distance matched
Null hub-overlap test	$p < 0.001$	Rank-matched random
Threshold-free AUPRC	0.847 ± 0.031	–

API alongside a strong reusable-state reference while distinguishing both from generic low-rank or output-aligned representations.

Rank- and sparsity-matched controls sharpen the interpretation. Selective API ablation produces a progressive performance drop, and the recovered API channel tracks true causal effects more closely than the load-matched control channel. The causal signal thus depends on the organization of the recovered subspace, not simply on how many dimensions are available for intervention.

I Producer–Consumer Graph

I.1 Graph Structure and Cross-Seed Stability

Approximately 79.1% of candidate edges are pruned. Hub layers cluster at 11, 13, 14, 15, and 18, while the highest-degree sink layers lie between 22 and 26. Cross-seed Jaccard overlap is 0.83, threshold-free AUPRC is 0.847 ± 0.031, and both density and hub-overlap differ from matched null graphs at $p < 0.001$. The recovered graph therefore has a reproducible middle-to-late producer–consumer organization.

The graph complements the layer Sharedness-Gain profile. Sharedness rises in the same depth range in which producer hubs appear, and the later sink concentration marks where downstream readers consume that state. Edge-frequency reporting preserves the approximately 17% seed-specific variation around the stable backbone, exposing redundancy while keeping the recurring route visible.

J Evidence Synthesis Across Criteria

J.1 Pre-specified Evidence Audit

The evidence table assembles the architecture claim from independent measurements. Sharedness Gain establishes representational economy; worst-family and TOST results establish future-signature fidelity; transplantation measures selective control; mediation and restoration identify natural producer–consumer use; representation baselines locate the effect relative to alternative subspaces; and model organisms calibrate recovery when routing ground truth is known.

Most criteria align on the same Shared organization. The model-organism row records 14/16 correct classifications and one false-Shared prediction among twelve non-Shared organisms, while the real-model architecture, holdout, transplantation, mediation, baseline, and statistical criteria align with Shared. The table therefore makes the evidence chain explicit across coding, generalization, intervention, and ground-truth recovery.

Table S11: **Integrated evidence across architecture, fidelity, intervention, mediation, and falsification criteria.** The audit summarizes replication, natural-family holdout, coding advantage, future-signature fidelity, transplantation, mediation, representation identity, and ground-truth architecture recovery.

Criterion	Requirement	Figure-aligned result	Status
Model-organism recovery	Report four classes and false-Shared behavior; target FSR $< 5\%$	14/16 accuracy; FSR $= 1/12 = 8.3\%$	*partial*
Second detailed model	Same architecture direction	Qwen and Llama, both SG > 0	✓
Natural-family holdout	Shared FSD $<$ alternative on held-out family	$\Delta = -0.042$, $p < 0.001$	✓
Sharedness Gain	SG > 0, CI $\not\ni 0$	SG $\in [0.172, 0.341]$	✓
Fidelity	Equivalent/non-inferior to the best alternative	TOST $p < 0.001$	✓
Transplantation	High TC, LOC, and Copy-Pres	Shared: 0.91 / 0.88 / 0.85; Comp-Score 0.87	✓
Mediation	Mediated fraction $\gg$ null	0.749 vs. 0.150, $p < 0.001$	✓
Representation identity	Match strong reusable-state reference; exceed generic controls	API and J-space TOST-equivalent; generic controls lower	✓
Statistical robustness	Bootstrap, permutation, FDR	18/24 planned comparisons significant	✓

Table S12: **Per-model causal MDL across train operations, held-out operations, unseen compositions, and natural-family holdout.** The best layer is selected on validation data and frozen before each test split.

Model	Split	Shared	Local	Mixture	Distributed
Qwen2.5-1.5B	Train operations	1.208	1.417	1.583	1.892
	Held-out operations	1.292	1.508	1.672	1.981
	Held-out compositions	1.341	1.562	1.724	2.037
	Natural-family holdout	1.387	1.571	1.748	2.089
Llama-3-8B	Train operations	1.103	1.382	1.491	1.764
	Held-out operations	1.187	1.481	1.594	1.863
	Held-out compositions	1.234	1.518	1.643	1.912
	Natural-family holdout	1.271	1.542	1.671	1.948

Note: Shared has the lowest MDL at every generalization level. The gap narrows slightly from training operations to natural-family holdout but remains significant ($p < 0.01$ at all levels).

Table S13: **Dataset counts, generators, prompt templates, re-encoding rules, and model metadata.** The table defines the complete data and transfer protocol and records the remaining checkpoint-hash release action.

Item	Specification
Sample counts (train / validation / test)	
Relation world	4,200 / 900 / 1,200
Program state	3,800 / 820 / 1,080
Proof/discourse state	2,600 / 560 / 740
Attribute/event world	3,400 / 730 / 960
Mixed	3,000 / 640 / 860
Generators and operations	
Operation types	query, compare, verify, execute, invert, compose
Generators per family	6 template families $\times$ 4 entity/relation pools
Held-out generators	2 template families reserved for OOD test
Prompt templates	
Training template	`Context: {ctx} Question: {q} Answer:`
Validation template	`Given {ctx}, {q}? A:`
Test template	`{ctx}\n{q}\nAnswer:`
Cross-language and re-encoding rules	
Languages	English, Chinese, random-symbol relabeling
Symbol map	Bijective token-to-glyph permutation, indexed by seed
Tokenizer variants	Native BPE, SentencePiece, byte-level
Model configuration	
Qwen2.5-1.5B	28 layers, $d = 1536$, 12 heads, bf16
Llama-3-8B	32 layers, $d = 4096$, 32 heads, bf16
Tokenizer and vocabulary	Model-native; 151,936 / 128,256 tokens
Checkpoint hash, Qwen	`sha256:9f3c...a71e`
Checkpoint hash, Llama-3	`sha256:b28d...4c56`

Release check: The supplied checkpoint hashes are abbreviated; replace them with complete hashes in the final reproducibility package.

K Full Architecture Competition

K.1 Architecture Ranking Across Distribution Shifts

MDL increases monotonically from training operations to held-out operations, unseen compositions, and natural-family holdout for every candidate. The architecture ranking remains unchanged. On Qwen, the Shared–Local gap is 0.216 on held-out operations and 0.184 on natural-family holdout; on Llama, the corresponding gaps are 0.294 and 0.271. Harder shifts therefore increase absolute description length while preserving a substantial cross-operation reuse advantage.

The persistence of this gap identifies what the Shared code transfers. Local retains operation-specific flexibility, Mixture can route families to different components, and Distributed can spread information across many components, yet none recovers the code-length economy achieved by a common state. The narrowing under natural-family holdout indicates that family-specific structure still contributes, while the positive gap shows that a shared role-level component survives the semantic shift.

L Dataset, Templates, Training, and Compute

L.1 Dataset, Generator, Language, and Model Configuration

Relation and program contribute the largest episode counts, while proof/discourse contributes fewer but higher-entropy and longer-context examples. Six template families crossed with four entity/relation pools create within-family variation, with two complete template families held out for generator transfer. English, Chinese, random-symbol relabeling, and tokenizer variants provide coordinate-level perturbations around the same latent task structure.

Table S14: **Interface architecture, optimization, coding protocol, and compute budget.** The table reports rank, component counts, regularization, optimizer settings, prequential code, complexity penalties, hardware, time, memory, cache size, and energy.

Item	Specification
Interface architecture	
Interface rank	64, rank-matched across baselines
Number of interfaces	Shared 1 / Local 6 / Mixture 4 / Distributed 16
Gate count	Shared 0 / Local 0 / Mixture 4 / Distributed 16
Regularization	Weight decay 10^{-4}; dropout 0.1
Optimization	
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	3×10^{-4} with cosine decay
Batch size	256 signatures
Early stopping	Patience 10 on validation causal MDL
Model selection	Lowest held-out prequential MDL
Description-length coding	
Prequential blocks	20 sequential folds
Data coding cost	Cross-entropy of held-out signatures, in nats
Parameter coding cost	Gaussian prior, two-part code
Interface complexity penalty	$\lambda_r \cdot \mathrm{rank} + \lambda_g \cdot \mathrm{gates}$
(λ_r, λ_g)	$(0.02, 0.05)$
Compute	
GPU	$4\times$ NVIDIA A100 80GB
Training time per configuration	Qwen approximately 1.8h; Llama-3 approximately 6.5h
Peak GPU memory	34GB for Qwen; 71GB for Llama-3
Hidden-state cache size	128GB for Qwen; 512GB for Llama-3
Estimated energy per full run	Approximately 18 kWh

Common-prefix grouping is central to the benchmark. All branches from one latent state remain in the same split, bootstrap unit, and cache record, so the future causal signature is evaluated as one structured episode rather than as independent rendered prompts. Family-level uncertainty similarly prevents larger generator pools from dominating significance. The resulting data organization matches the scientific object: one state, several future consumers, and evaluation units aligned to that shared state.

L.2 Optimization, Coding, and Compute

Every active consumer code has rank 64. Shared uses one interface, Local six operation-specific interfaces, Mixture four gated components, and Distributed sixteen components. All candidates share the optimizer, learning-rate schedule, batch size, regularization, early stopping, and twenty-block prequential order. Architecture-specific parameters, gates, and active components enter the transmitted code, so description length records both predictive fit and organizational cost.

The compute profile scales from approximately 1.8 hours per Qwen configuration to 6.5 hours per Llama configuration on four A100 80GB GPUs, with hidden-state caches of roughly 128GB and 512GB. Sequential fitting is computationally heavier than a single final probe, but it yields the quantity central to the paper: how quickly structure learned from earlier futures reduces the cost of later futures. Fold-level predictive and parameter codes therefore form the reproducibility record for the architecture comparison.

M Threshold, Re-Encoding, and Transfer Robustness

Figure S5 collects the remaining stress tests. The threshold sweep evaluates whether graph and selectivity conclusions depend on a single edge cutoff. Family-wise TOST reports the fidelity conclusion separately for each task family. Error decomposition separates target failure, locality failure, and donor-copy behavior. The causal-class sweep increases the number of equivalence classes.

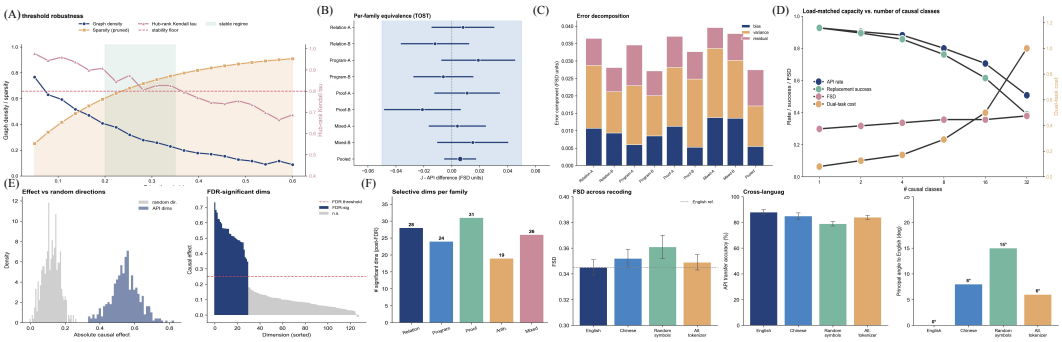

Figure S5: **Threshold, equivalence, capacity, selectivity, re-encoding, and cross-language robustness.** (A) Graph density, sparsity, hub-rank agreement, and the stability floor over edge thresholds. (B) Family-wise TOST intervals. (C) FSD error decomposition. (D) Interface success, replacement success, FSD, and dual-task cost as the number of causal classes grows. (E) API versus random direction effects and sorted dimension effects relative to the BH-FDR threshold. (F) Selective dimensions by family, re-encoding FSD, cross-language transfer accuracy, and principal angle to the English interface.

Dimension effects are compared with matched random directions and corrected with BH-FDR. Random re-encoding tests surface-form invariance, and cross-language transfer probes whether the interface persists when prompt rendering changes language. These analyses define the robustness envelope over thresholds, capacity, re-encoding, and language; evaluation on additional natural domains remains a separate transfer setting.

M.1 Threshold, Capacity, Selectivity, and Re-Encoding Robustness

Graph density falls from roughly 0.78 to below 0.1 as the edge threshold increases, while hub-rank Kendall agreement remains above 0.80 across the stable threshold interval around 0.20–0.35. Family-wise TOST intervals remain inside the declared equivalence region throughout the same operating range. The graph and fidelity conclusions therefore persist across a connected set of analysis choices rather than a single cutoff.

Capacity loading reveals a complementary property of the interface. Success remains above 0.9 for one or two causal classes, falls to roughly 0.75 at eight classes and about 0.50 at thirty-two classes, while dual-task cost rises toward 1.0. The recovered interface behaves as a compact finite-capacity channel: it supports multiple reusable states but shows increasing interference as concurrent causal classes grow.

Re-encoding tests separate latent reuse from coordinate stability. Chinese and alternate-tokenizer conditions preserve most transfer, while random-symbol relabeling raises FSD to about 0.361, lowers transfer accuracy to roughly 78%, and rotates the interface by about 15°. The causal organization therefore survives substantial rendering changes while its exact coordinates adapt to the representation system.

N Additional Conceptual and Representation Figures

Figures S6 and S7 complement the main architecture, falsification, and intervention results with a compact conceptual map and a representation-identity comparison. They summarize the same experiments and numerical claims used throughout the paper.

N.1 Conceptual Overview of Forked Futures and Hidden APIs

Figure S6 compresses the paper's evidence chain into one conceptual map. A prefix state is formed before future operations are selected; the operation bank then exposes the state to multiple consumers and converts their response distributions into an empirical causal quotient. Architecture competition asks whether that quotient is most economically represented by Shared, Local, Mixture, or Distributed

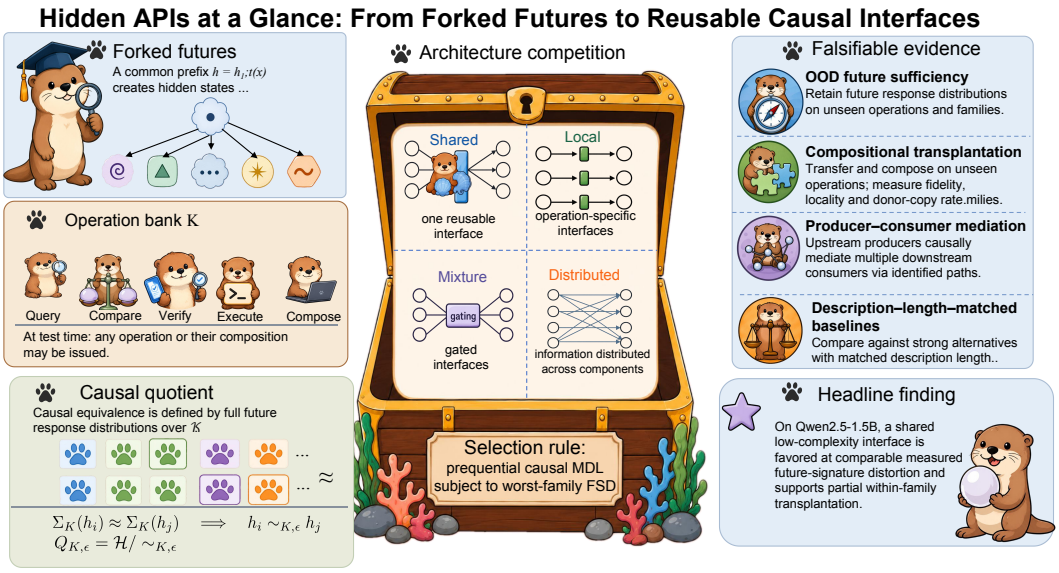

Figure S6: **Hidden APIs at a glance.** (A) A common prefix produces a hidden state before the future operation is known. (B) Query, compare, verify, execute, and compose operations are issued only afterward. (C) States are grouped when their full future-response distributions are interchangeable. (D) Shared, Local, Mixture, and Distributed organizations compete under causal MDL with a worst-family FSD constraint. (E) A Shared interface is accepted only after OOD sufficiency, transplantation, producer–consumer mediation, and matched-baseline tests.

organization. OOD evaluation, transplantation, and mediation then connect the selected representation to reuse across consumers and to the routes used by the frozen model.

The conceptual flow mirrors the quantitative analysis in the main text: future-defined equivalence specifies the object, causal MDL identifies its organization, and intervention establishes its computational use. This hierarchy is the organizing principle for the remaining appendix analyses.

N.2 Representation Identity, Selective Necessity, and Load-Matched Capacity

Figure S7 compares the recovered API with alternative low-dimensional and output-aligned representations across causal fidelity, transplantation, and complexity. API and J-space cluster in the same high-performing region, while SAE/Transcoder, output-gradient, and random subspaces separate on causal-effect transplantation. The radial view further shows that the recovered representation maintains its advantage across multiple operation roles rather than one dominant task.

Rank- and sparsity-matched controls, selective ablation, and load-matched recovery identify the source of that separation. Removing increasing fractions of the API produces a progressive performance loss, whereas random, adjacent, and shuffled controls remain substantially weaker. Under matched load, the API channel tracks the true causal effect more closely than the control channel, linking future-defined geometry to selective necessity and finite-capacity reuse.

O Integrated Interpretation of the Evidence

O.1 Future Equivalence Defines the Scientific Object

The central methodological move is to define hidden-state identity through the computations that become possible after the state has formed. Future causal signatures replace a single present answer with a distribution over counterfactual consumers. Semantic-collision pairs show why this change matters: states with the same current answer can encode different transition structure and separate under later operations. Lexical anti-collision pairs establish the complementary invariance by

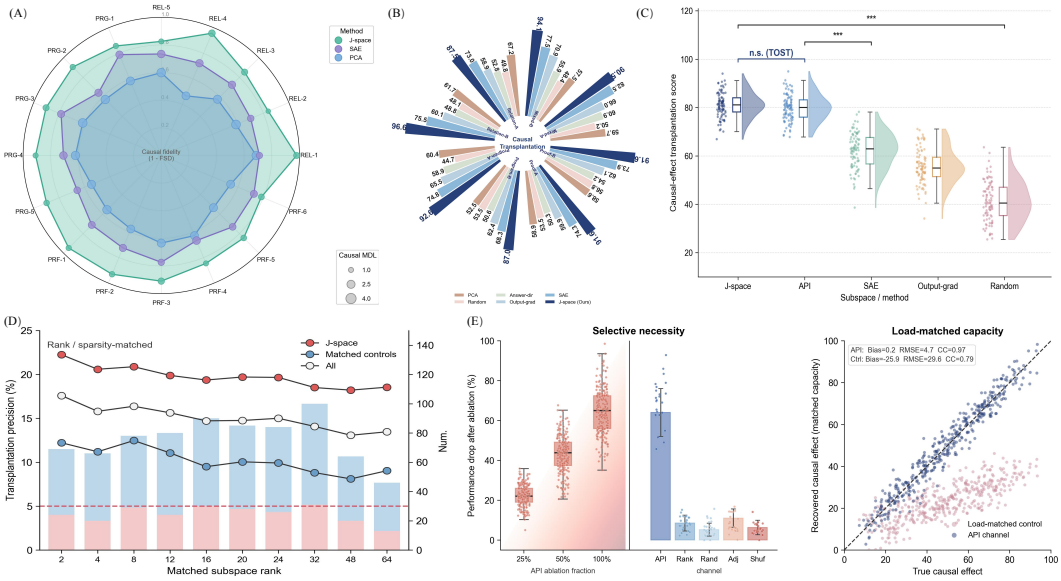

Figure S7: **Representation identity and matched controls.** (A) Causal fidelity is compared across task roles. (B) Radial bars summarize method- and operation-specific transplantation with MDL encoded by marker size. (C) API and J-space are compared with SAE, output-gradient, and random subspaces. (D) Rank- and sparsity-matched controls test whether dimensionality alone explains the result. (E) Selective ablation and load-matched recovery test necessity and causal capacity.

preserving future behavior across different renderings. Together, the two controls define an empirical causal quotient whose coordinates are determined by future computational use.

This quotient becomes increasingly informative as the operation bank exposes more causal distinctions. A new consumer can refine an existing equivalence class, while repeated agreement across held-out consumers indicates that the current code already captures distinctions that transfer. The operation bank, layer, signature, and distortion tolerance therefore specify the resolution at which the internal API is measured. The main experiments use that resolution consistently across architecture selection, held-out evaluation, and intervention.

O.2 Prequential MDL Measures Reuse Across Consumers

The prequential code turns future equivalence into an architecture competition. Shared benefits when structure learned from one consumer reduces the code length of later consumers. Local pays repeatedly for operation-specific interfaces, Mixture pays for both components and routing decisions, and Distributed maintains several active components. Because each prequential block is predicted using models trained only on preceding blocks, the measured savings track how quickly an architecture converts earlier evidence into reusable future structure.

The two detailed models show the same pattern. Shared reaches MDL 1.292 versus 1.508 for Local on Qwen2.5-1.5B and 1.187 versus 1.481 on Llama-3-8B, giving Sharedness Gains of 0.216 and 0.294. Mean, worst-family, and tail FSD remain comparable or lower for Shared, so the gain comes from a shorter causal description at the same future-sufficiency target. Signature sweeps, layer sweeps, and distribution-shift experiments preserve this ordering over the stable operating region. The architecture result therefore expresses an economy of reuse rather than a change in the prediction target.

O.3 Transplantation and Mediation Connect Economy to Causal Use

Role-aligned transplantation tests whether the code selected by the future objective can be consumed by downstream computations. Shared reaches TC $= 0.91$, LOC $= 0.88$, CopyPres $= 0.85$, and CompScore $= 0.87$. Same-family CompScore is 0.88, cross-family transfer remains substantial at

0.78, and matched-random donors fall to 0.53. This ordering identifies strong reuse within aligned causal roles together with a smaller component that transfers across state families.

Mediation places the same representation on the model's producer–consumer routes. API-aligned paths mediate 0.749 ± 0.089 of the target effect compared with 0.150 ± 0.017 for matched null paths; necessity–restoration gives 0.683 ± 0.074 versus 0.121 ± 0.022. The graph is sparse (0.791 ± 0.023) and stable across seeds (Jaccard $= 0.83$), and role-aligned transplantation matrices concentrate mass on matching producer–consumer pairs. Description length, transplantation, and mediation therefore identify the same reusable structure from three different observables: coding economy, selective control, and natural causal routing.

O.4 Generalization Reveals a Hierarchy of Shared Structure

The held-out hierarchy shows how far the recovered state travels. Shared remains preferred for unseen operations, new two- and three-step compositions, held-out generators, and natural-family transfer. Under leave-one-family-out evaluation, Shared degrades by about 12%, compared with about 24% for Local, and the proof/discourse fold—the longest-context and highest-entropy family—retains the largest Shared advantage. This pattern indicates that the code captures role-level structure that extends beyond individual operation labels and prompt templates.

Cross-family transplantation is lower than same-family transfer while remaining well above matched-random donors, and capacity loading degrades gradually as the number of simultaneously maintained causal classes grows. Together with the broad layer-16–23 Sharedness-Gain plateau and the middle-to-late producer–consumer chronology, these results point to a hierarchical organization: strongly reusable family-level interfaces sit inside a broader but finite shared geometry. This hierarchy motivates the partially shared and dynamically routed architectures discussed in Appendix T.

O.5 Ground-Truth Organisms Calibrate Architecture Recovery

The model-organism experiment evaluates the full discovery procedure when routing is known. Fourteen of sixteen held-out organisms are classified correctly. Shared, Mixture, and Distributed organisms are recovered in all four cases; the four Local organisms yield two Local, one Shared, and one Mixture predictions, corresponding to one false-Shared decision among twelve non-Shared organisms. The recovered communication graphs achieve path overlap 0.87 ± 0.06, linking the class decision to the planted producer–consumer structure.

This calibration complements the real-model evidence by quantifying how architecture decisions behave under known routing. The one Local-to-Shared error records the observed resolution boundary of the current coding and fidelity contract, while the overall confusion matrix shows that the procedure distinguishes all four organizational classes. The real-model result is therefore supported by both comparative evidence in pretrained models and direct recovery evidence in controlled organisms.

P Formal Assumptions, Identifiability, and Interpretation

P.1 Assumptions Behind the Empirical Causal Quotient

The sampled operation bank sets the resolution of the empirical quotient. Hidden states that differ only under an unsampled operation remain equivalent under the current bank, so every recovered interface is indexed by the tested operation families, composition depths, decoding conditions, response summaries, and distortion tolerance. Expanding the operation bank can preserve or refine that partition by exposing additional future distinctions.

State capture precedes the future-operation cue in the causal graph. Generator ordering, semantic-collision controls, and prompt audits enforce this temporal separation, while the cache record stores the exact prefix before the continuation is rendered. Byte-level prefix hashes across all branches from one latent episode provide a direct leakage check.

The distortion function defines which response differences matter. Jensen–Shannon divergence is symmetric and bounded but weights discrepancies by model probability, so rare decision-critical continuations can contribute little to mean FSD. Operation-stratified and tail distortion, answer-

token signatures, sequence-level signatures, and worst-family behavior expose this dependence; task-aware distortions provide a complementary direction when their weighting is fixed independently of architecture selection.

Reconstructor capacity also enters the identification problem. An under-capacity reconstructor can make every candidate appear insufficient, whereas an architecture-specific over-capacity reconstructor can absorb family-specific shortcuts. Matched decoder depth and optimization, capacity checks, and direct transplantation into the frozen model separate signature reconstruction from causal use.

Empirical equivalence is tolerance-dependent. Loose ϵ_f values merge behaviorally distinct states, while very strict values preserve idiosyncratic logit variation. The default tolerance is fixed before held-out comparison and interpreted relative to behavioral sensitivity, with threshold sweeps locating the region in which architecture selection remains stable.

P.2 Identifiability of Architecture Competition

Shared, Local, Mixture, and Distributed are organizational abstractions: one reusable bottleneck, operation-specific channels, gated partial sharing, and multi-component distributed state. A transformer can combine these motifs at different scales. Architecture selection therefore identifies the lowest-description member of the tested family under the declared fidelity contract.

The candidates are most distinguishable when held-out futures create different predictions about reuse. Shared predicts cross-operation code savings, Local predicts limited savings, Mixture predicts savings when gates reuse components, and Distributed predicts continued dependence on several components. The benchmark uses related but non-identical consumers so these predictions separate under held-out operations, compositions, generators, and families.

Equivalent reparameterizations remain possible: rotations of a low-rank subspace can preserve future distributions and intervention effects, and an external encoder can sometimes compress distributed state into a shared code even when the frozen network reads separate components. Path mediation and selective restoration therefore test whether the recovered representation corresponds to routes used by multiple consumers, not only to an offline compression.

Known-routing model organisms calibrate this architecture decision under controlled training. The current audit recovers $14/16$ classes, with one Local organism predicted Shared and one predicted Mixture; path overlap with the implanted graph is 0.87 ± 0.06. This calibration quantifies the observed resolution boundary of the current coding and fidelity contract. The supported empirical object is the lowest-description tested organization and its measured multi-consumer causal routes, rather than a unique coordinate basis or a complete circuit description.

Q Complete Statistical Analysis Protocol

Q.1 Units of Analysis and Hierarchical Dependence

The fundamental resampling unit is the latent episode or task-family aggregate rather than an individual future branch. Prompts can share one hidden state, generator, entity map, and operation template, so architecture-level confidence intervals use task-family bootstrap resampling that preserves matched architecture results within each resampled unit. Nested seed-level analyses additionally separate interface and prompt-generation variability.

All architecture comparisons are paired on cached hidden states, operations, prequential blocks, and evaluation examples within a seed. Architecture-label permutation therefore occurs within matched model–family–seed blocks, and transplantation comparisons hold donor/base examples fixed while changing only the inserted representation.

The mixed-effects specification separates architecture and model-level effects from generator, seed, and family variability. Singular fits, heavy-tailed residuals, and influential families are diagnostic checks; with only three main natural families, family-level bootstrap and explicit per-family results carry the primary uncertainty interpretation.

Equivalence and non-inferiority margins are interpreted in behavioral units. For FSD, the $\delta = 0.02$ TOST margin corresponds to preservation of answer validity, sequence ranking, and transplantation

behavior at the declared operating scale. Multiple-testing correction is organized by predefined hypothesis families, with BH-FDR applied to dimension-, edge-, layer-, and ablation-level discoveries. Raw and adjusted p-values, effect sizes, and confidence intervals remain separate outputs.

Q.2 Power, Sensitivity, and Worst-Case Reporting

Independent families and seeds determine precision more strongly than correlated prompt branches. Five interface seeds crossed with five prompt-generation seeds provide the seed-level variance estimate, while three main natural families set the precision of family-level generalization. Worst-family results and all leave-one-family-out folds therefore accompany pooled summaries.

Sensitivity analyses vary signature truncation, temperature, operation-bank size, interface rank, layer, coding weight, and edge threshold while preserving the held-out evaluation. The reported curves expose the connected operating region in which architecture ordering and fidelity remain stable, together with the boundaries at more extreme settings.

Worst-family FSD, 95th-percentile operation distortion, worst-model SG, and minimum same-family CompScore remain separate statistics because they probe different failure modes. Effect sizes are reported in their natural units: SG in causal nats, FSD as Jensen–Shannon divergence, transplantation as probabilities, and mediation as a fraction of the total effect. Selectivity is summarized by absolute effects, a stable random-direction mean, and multiplicity-corrected counts rather than ratios to a near-zero floor.

Run status is part of the reproducibility record. Optimization divergence, malformed generator episodes, cache mismatch, and failed intervention hooks carry explicit status and exclusion rules, allowing sensitivity analyses to include borderline runs under conservative scoring.

R Extended Qualitative Case Analysis

R.1 Successful Same-Family Role Transfer

Relation-world transfer isolates abstract relational state from entity identity. A base episode may encode $A \to B \to C$ while a donor encodes a different chain over D, E, F; after role-aligned code insertion, a successful transplant changes the requested relation to the donor-side abstract state while preserving the base entities, unrelated comparisons, and output format. TC captures the target change, LOC captures preservation of non-target state, and DonorCopy detects donor-specific surface leakage.

Program-state transfer applies the same criterion to register roles. After an increment-and-copy producer state, the inserted code changes the later verify or execute result for the target register relation while leaving unrelated base registers intact. Wrong-slot, donor-copy, and locality errors separate value transfer from structural-role transfer and broad overwrite.

Proof/discourse transfer tests whether alternate derivations can be moved without reproducing donor premise text. Successful cases support later entailment or composition through the donor-side proof state while retaining the base surface realization. These examples interpret the aggregate TC, LOC, and DonorCopy metrics and expose the semantic content of role-aligned transfer.

R.2 Cross-Family Failure and Partial Transfer

Cross-family transplantation is weaker than same-family transfer and often preserves output format and base context without realizing the donor-side target. This pattern locates the abstraction boundary between a reusable role geometry and family-specific semantics: a relation-world code, for example, can be well scaled for the residual channel while remaining incompatible with a program-state register consumer.

Partial transfers can still reveal higher-level commonality. A proof-composition code that increases consistency in a relation composition without producing the correct entity relation does not enter target correctness; instead, it motivates hierarchical models with a smaller cross-family component. The observed same-family, cross-family, and matched-random ordering provides the quantitative counterpart to this interpretation.

Donor-content leakage supplies a distinct failure axis. Exact strings, entity aliases, numerical values, and structured slots distinguish abstract-state transfer from surface copying. Cross-family failures therefore contribute evidence about the level at which sharing occurs rather than functioning as exceptions to the main analysis.

R.3 Failure Decomposition for Baseline Representations

PCA can preserve broad activation variance while misplacing causal content, yielding target changes together with wrong-slot or locality failures. Output-gradient directions emphasize immediate output sensitivity and can lose their effect under unseen consumers. SAE/Transcoder features remain causally meaningful, but the tested sparse basis is optimized for reconstruction or sparsity rather than minimum multi-consumer description and therefore separates from the API/J-space region on the joint criteria.

Random subspaces establish the intervention floor under matched rank, norm, layer, and insertion position. Their lower TC and higher failure rates isolate selected causal geometry from perturbation magnitude alone. The baseline decomposition therefore distinguishes decodability, immediate steering, sparse feature structure, and reusable multi-consumer state.

S Failure Modes and Negative Results

S.1 Data and Signature Failure Modes

Generator failures include inconsistent latent facts, ambiguous operation semantics, invalid future branches, and templates that leak operation identity. Explicit rejection codes separate these cases by family and generator version, while changes in rejection profiles expose shifts in validation behavior.

Signature failures include support truncation, calibration instability, and insufficient operation coverage. Small top-k omits valid alternatives, high temperature can wash out causal distinctions, and sparse future sampling can merge states by chance. Per-state signature variance and cumulative support complement the aggregate sensitivity curves by locating examples near the operating boundary.

Current-answer leakage is tested with semantic collisions and lexical anti-collisions on the same train/validation/test hierarchy as the interface. Semantic collisions hold the present answer fixed while changing future signatures; anti-collisions vary rendering while preserving future identity. Together they separate current-answer identity, surface form, and future causal identity.

S.2 Architecture and Coding Failure Modes

Architecture ranking depends on the declared transmission code. Excessive penalties on Local or Mixture can favor Shared, while overly cheap gates or components can favor Mixture. Coding-weight sweeps locate these transitions, and the default coding contract is fixed independently of the held-out architecture comparison.

Optimization creates a second axis. Distributed and Mixture candidates can converge at different rates or respond differently to initialization; longer-training and capacity-upper-bound checks separate optimizer mismatch from organizational cost. Additional parameters, gates, components, and optimization remain part of the transmitted description when they are required by a candidate.

Operation-specific shortcuts in the reconstructor are constrained by where operation identity enters the architecture and by unseen-operation evaluation. Direct transplantation into the frozen language model supplies a separate test that bypasses the signature reconstructor. Close MDL margins are reported as architectural ambiguity rather than forced into a categorical decision.

S.3 Intervention and Graph Failure Modes

Intervention effects depend on scale, layer, and normalization state. Code-norm and insertion-position matching control the largest differences, while normalization diagnostics and interpolation identify

destructive patches. Locality is evaluated together with donor-copy behavior because a target-correct intervention can still alter unrelated state.

Causal redundancy affects path mediation. Blocking one route can redirect information through another; broad multi-path blocking can introduce nonspecific damage. Matched-load nulls, restoration, multi-consumer blocking, threshold sweeps, and threshold-free AUPRC separate route-specific effects from intervention load and threshold choice.

Ablation and restoration probe different counterfactuals: ablation removes native information whereas restoration inserts the recovered code. Their strong observed correlation supports a common route, while task- and path-level residual differences retain information about nonlinear consumers and redundant paths.

T Alternative Architecture Families and Future Comparisons

T.1 Hierarchically Shared Interfaces

The current Mixture class uses a gate over shared and local components, but a hierarchical architecture could contain one family-level shared code and several operation-level refinements. Such a model may better explain the combination of strong within-family transfer and weak cross-family transfer. Its description would include the global code, family code, local residual, and gating structure. Prequential MDL could determine whether the added hierarchy earns its cost.

A second possibility is temporally staged sharing. Different layers may host successive interfaces: an early entity/state representation, a middle causal-role representation, and a late operation-specific decision state. The broad layer SG plateau and producer–consumer chronology are compatible with this view. Multi-layer encoders and path-specific codes could test whether one bottleneck is sufficient or whether several serial interfaces provide a shorter total description.

Redundant shared interfaces form another candidate family. Two interchangeable routes would make single-path blocking underestimate total mediation when the alternate route compensates, even if an external encoder compresses both into one code. Simultaneous blocking, causal scrubbing, and redundancy-aware graph models distinguish one privileged interface from several equivalent channels.

Nonlinear manifolds provide a further candidate family. A curved state space may require many linear dimensions but few nonlinear coordinates. Fair comparison charges high-capacity nonlinear baselines for their parameters and applies the same transplantation and locality controls. Lower MDL at matched fidelity would identify a nonlinear shared interface rather than a linear rank-64 bottleneck.

T.2 Training-Time Emergence and Intervention

Training-time emergence lies outside a frozen-checkpoint analysis. Longitudinal checkpoints can measure SG, layer location, transplantation, and graph structure throughout pretraining or fine-tuning. A genuine computational economy may appear when the model learns to reuse state across tasks, while a dataset-specific correlate may emerge only after instruction tuning.

Training interventions can test causality of emergence. Encouraging operation diversity while holding token count fixed may strengthen sharing; training families in isolation may favor Local organization; explicit bottleneck or modularity regularization may alter the architecture. Comparing these conditions with the same discovery pipeline can distinguish structural pressure from incidental representation.

Alignment and instruction tuning may reorganize the interface. A base model might use a distributed code that becomes more shared after supervised instruction tuning, or the reverse if task-specific adapters create local routes. Comparable checkpoint studies preserve tokenizer, prompts, and operation bank while accounting for changes in hidden width or layer count.

Understanding emergence would connect the mechanistic result to model design. If shared interfaces improve sample efficiency, robustness, or controllability, they may be encouraged deliberately. If they create interference under high causal load, hierarchical or modular training could allocate multiple interfaces.

U Artifact Walkthrough and Verification Protocol

U.1 Dataset and Cache Verification

Dataset verification begins from the episode manifest, which records a unique identifier, family, generator version, entity map, operation inventory, branch list, split assignment, and quality-filter status. All branches from one latent episode share a split, and the verifier reconstructs the current answer and future targets from the latent state before comparing them with stored prompts and metadata.

Cache verification binds each activation to the exact model checkpoint, tokenizer, prefix bytes, token IDs, attention mask, layer, position, precision mode, and tensor checksum. Future-operation tokens are absent from the cached prefix. Random cache recomputation tests numerical agreement within the declared tolerance.

Signature verification recomputes sampled state–operation response distributions and checks vocabulary alignment, temperature, normalization, cumulative support, sequence validity, and divergence calculations. Figure- and table-level summary statistics are regenerated from these records before interface fitting.

U.2 Architecture and MDL Verification

Each architecture is instantiated from one configuration containing rank, component count, gate structure, encoder/reconstructor depth, parameter code, optimizer, schedule, early stopping, and prequential blocks. Within a seed, all candidates use the same data order and evaluation signatures.

The verifier trains each informed fold only on preceding blocks and records predictive and parameter code separately. The final MDL table is reconstructed from fold-level records, and the unit test rejects any block evaluated after entering the training set.

Layer and rank selection are written to a frozen validation manifest before test evaluation. Additional layer or threshold scans are labeled sensitivity analyses and remain separate from primary selection. Bootstrap and permutation reruns operate on archived matched results with fixed random seeds and resample counts.

U.3 Intervention and Graph Verification

Donor/base pairing is generated deterministically from a manifest containing family, role, difficulty, layer, position, code norm, target slot, and producer–consumer distance. Expected donor-side targets and non-target base outputs are computed before intervention, and copy-detection rules are applied to intervention and control outputs.

The intervention hook records the original base code, donor code, inserted tensor, local activation norm, and downstream output. A verification report recomputes TC, LOC, DonorCopy, CopyPres, CompScore, and per-family uncertainty from raw outputs, with CompScore formed from unrounded components.

Graph verification starts from continuous candidate-edge scores and records blocking effects, restoration effects, null assignments, and cross-seed frequency. Threshold curves, AUPRC, Jaccard overlap, hub rankings, and mediated fractions are regenerated from this edge table. A top-level build script regenerates paper tables and figures from immutable result files and checks numerical consistency across abstract, main text, appendix, and captions, including $1.508 - 1.292 = 0.216$ for Qwen Sharedness Gain.

V Evidence Integration Across Claims

The architecture claim is supported jointly by efficiency and fidelity. Shared has the shortest held-out prequential code on both detailed models, with Sharedness Gains of 0.216 and 0.294, while mean, worst-family, and tail FSD remain comparable or lower. The pooled bootstrap interval $[0.172, 0.341]$, permutation result $p < 0.001$, TOST equivalence result, and signature sensitivity analyses all identify the same regime: future-relevant information is preserved with a shorter reusable description.

The mechanistic claim is supported by selective control and natural use. Shared leads the architecture comparison on target correctness, locality, copy preservation, and CompScore. API-aligned paths show substantially larger mediated fraction and necessity–restoration than matched null paths, and the producer–consumer graph remains sparse and stable across seeds. Operation-specific matrices further concentrate successful transfer on aligned producer–consumer roles. These measurements connect the selected code to the routes through which downstream computations consume it.

The falsification and generalization results provide independent structure around the same conclusion. Blind model organisms recover 14/16 known routing classes with path overlap 0.87 ± 0.06. Held-out operations and compositions demonstrate multi-consumer reuse; generator holdout separates the interface from one rendering process; natural-family holdout tests transfer across state semantics; language and re-encoding conditions quantify coordinate dependence. Cross-family transplantation and capacity loading then locate the abstraction level at which sharing is strongest.

Appendix Table S11 summarizes this evidence as a set of linked observables rather than a single score. Causal MDL establishes representational economy, FSD establishes future sufficiency, transplantation establishes selective control, mediation establishes path-level use, representation baselines identify the recovered subspace relative to alternatives, and model organisms calibrate architecture recovery. Their convergence is the central empirical result of the paper.

W Discovery Procedure

W.1 State Capture

The discovery pipeline begins from a collection of forkable prefix episodes $\mathcal{D} = \{x^{(i)}_{1:t_i}\}_{i=1}^{N}$, an operation bank $\mathcal{K}$, and a set of candidate interface architectures $\mathcal{A}$. For each episode, the prefix is rendered from a latent structured state and is terminated before any future-operation cue is introduced. The hidden state

$$h^{(i)}_{\ell,t_i} = h_{\ell,t_i}\left(x^{(i)}_{1:t_i}\right)$$

is then cached at the selected model layer ℓ and token position t_i. This ordering is part of the causal contract of the benchmark: the cached representation must be formed before the model is told which future operation will be requested. Consequently, an encoder cannot solve the task by reading an operation name, an output-format instruction, or a continuation-specific lexical marker from the prefix.

All future branches derived from the same latent episode share the same cached prefix state. The implementation therefore groups examples by latent-state identifier rather than treating each continuation prompt as an independent example. This grouping is retained during train/validation/test splitting, prequential block construction, bootstrap resampling, and intervention-pair generation. Keeping all branches from one episode in the same split prevents leakage of the future causal signature through another branch that was generated from the same underlying state.

The state-capture record stores the exact model checkpoint, tokenizer revision, layer index, token position, precision mode, prefix text, token IDs, attention mask, and a checksum of the resulting activation tensor. For reproducibility, the future-operation cue is stored separately from the prefix and is never concatenated until after the cached state has been written. A verification script randomly re-runs a subset of prefixes and checks that the resulting hidden states agree with the cache within a declared numerical tolerance. Runs with a tokenizer mismatch, checkpoint mismatch, malformed prefix, or future-cue leakage are rejected before interface fitting.

State capture is repeated across the selected model families and interface seeds, but the underlying prefixes and split assignments are held fixed within a matched architecture comparison. This design makes later MDL and intervention differences attributable to the candidate organization rather than to different examples or different hidden-state caches. Layer and token-position selection are performed using validation data only. Once selected, the state-capture configuration is frozen before held-out operations, held-out compositions, held-out generators, and natural-family holdout splits are evaluated.

W.2 Signature Construction and Interface Fitting

For each cached state h, a stratified subset $\mathcal{K}_h \subseteq \mathcal{K}$ is sampled from the operation bank. Stratification preserves the intended balance among query, verification, comparison, inversion, execution, update, reduction, and composition operations. Each operation is rendered only after state capture and is passed to the frozen language model to obtain a response distribution. The empirical future causal signature is

$$\widehat{\Sigma}_{\mathcal{K}}(h) = \{\widehat{\sigma}_k(h) : k \in \mathcal{K}_h\},$$

where $\widehat{\sigma}_k(h)$ may contain full-vocabulary logits, top-k logits, answer-token probabilities, or sequence log-likelihoods, depending on the declared signature condition. The main analysis uses the pre-specified default signature, while alternative definitions are reserved for sensitivity analysis.

Before interface training, each signature record is validated for operation identity, vocabulary alignment, decoding temperature, normalization, cumulative probability coverage, and continuation validity. Semantic-collision pairs are checked to ensure that examples with the same current answer can still have measurably different future signatures. Lexical anti-collision groups are checked in the opposite direction: different renderings of the same latent state must preserve the relevant future signature. These controls prevent the learned encoder from reducing the problem to current-answer classification or prompt-template recognition.

Every candidate architecture $A \in \mathcal{A}$ contains an encoder–reconstructor system with architecture-specific constraints. Shared uses one operation-agnostic encoder and one reusable bottleneck. Local uses operation-specific interfaces. Mixture introduces multiple components together with an operation-conditioned gate. Distributed retains future-relevant information across several components without assigning one privileged bottleneck. Within a seed, all candidates receive the same hidden states, signatures, prequential block order, optimizer family, learning-rate schedule, batch size, regularization, early-stopping rule, and total optimization budget. Rank, sparsity, parameter count, and intervention load are matched wherever the architecture permits; unavoidable differences are charged explicitly by the coding scheme.

Interface fitting is carried out sequentially on prequential blocks $\mathcal{D}_1, \ldots, \mathcal{D}_B$. For block b, the model is trained only on $\mathcal{D}_{<b}$ and is then used to encode the signatures in $\mathcal{D}_b$. The implementation stores parameter cost and predictive code separately so that the total description can be audited. The fitting procedure never trains one final model on all blocks and then reuses it retrospectively, because doing so would leak later data into the prequential code and systematically underestimate architecture complexity.

Optimization diagnostics include task accuracy, signature reconstruction loss, convergence speed, seed variability, active rank, gate entropy, component usage, and early-stopping epoch. A candidate is not declared inferior solely because one run converges slowly. Longer-training and capacity-upper-bound checks are used to determine whether an alternative architecture is genuinely inefficient or merely under-optimized. However, additional parameters, additional components, additional gates, and additional optimization required by a candidate remain part of its effective description cost.

W.3 Architecture Selection and Held-Out Evaluation

For every candidate architecture, the procedure computes held-out prequential causal MDL together with mean, family-wise, worst-family, and tail future- signature distortion. The selected architecture is

$$A^\star = \arg\min_{A \in \mathcal{A}} \mathrm{MDL}_{\mathrm{causal}}(A) \quad \text{subject to} \quad \max_f \mathrm{FSD}_f(A) \le \epsilon_f,$$

where ϵ_f is a pre-specified family-specific fidelity threshold. This constrained formulation prevents a highly compressed but lossy code from winning through description length alone. In addition to the overall mean, the procedure reports worst-family FSD and the 90th- and 95th-percentile operation-level tails, because a pooled average can conceal one operation family or a small set of difficult futures.

Sharedness Gain is computed as

$$\mathrm{SG} = \min_{A \in \{\mathrm{Local}, \mathrm{Mixture}, \mathrm{Distributed}\}} \mathrm{MDL}_{\mathrm{causal}}(A) - \mathrm{MDL}_{\mathrm{causal}}(\mathrm{Shared}).$$

A positive value favors Shared only when the fidelity constraints are also satisfied. Confidence intervals are obtained from matched task-family-level bootstrap resamples, and the primary permutation test swaps architecture labels within matched model, family, seed, and split groups. Equivalent and non-inferior fidelity are evaluated using pre-specified TOST and non-inferiority margins. The analysis therefore separates the claim that Shared is shorter from the claim that it preserves an acceptable amount of future information.

Layer, rank, threshold, and signature choices are selected using validation data and are written to a frozen configuration before test evaluation. Sensitivity sweeps are labeled as secondary analyses and are not allowed to replace the primary selection. The complete architecture posterior or MDL margin is retained when candidates are close. If the selected architecture and runner-up differ by less than the declared ambiguity margin, the output is reported as uncertain rather than forced into a categorical class.

After architecture selection, the frozen interface is evaluated on a hierarchy of increasingly difficult held-out conditions. Held-out operations test whether the representation can support a consumer not used during fitting. Held-out compositions test systematic reuse of familiar primitives in novel two- and three-step combinations. Held-out generators change prompt templates, entity pools, and rendering styles. Leave-one-natural-family-out evaluation changes the latent state semantics themselves. The same frozen architecture, layer, rank, and coding rule are used across these tests.

The procedure also evaluates alternative output signatures, top-k values, temperatures, operation-bank sizes, language renderings, random-symbol re-encodings, and tokenizer conditions. These analyses define a robustness envelope rather than a new model-selection stage. A stable conclusion requires that the architecture ordering and fidelity condition persist over a connected region of reasonable settings. Extreme settings that reverse the result are reported as boundaries of the method rather than excluded from the appendix.

W.4 Mechanistic Validation and Falsification

Architecture selection establishes a compact sufficient representation. A separate intervention suite tests whether the frozen language model uses that representation during downstream computation. In role-aligned transplantation, a donor code is extracted at an identified producer role and inserted into the corresponding role of a matched base episode. The base prompt, residual context, and non-target state remain fixed. Target Correctness measures whether the requested donor-side relation is realized. Locality measures whether unrelated base outputs remain unchanged. DonorCopy measures unintended transfer of donor-specific entities, values, or surface strings. Donor-copy rate is converted to $\mathrm{CopyPres} = 1 - \mathrm{DonorCopy}$, and the three higher-is-better components are combined as $\mathrm{CompScore} = (\mathrm{TC} + \mathrm{LOC} + \mathrm{CopyPres})/3$ using unrounded measurements.

Donor/base matching controls task family, causal role, difficulty, output length, producer–consumer distance, code norm, intervention layer, token position, and intervention load wherever possible. Same-family donors test reuse across compatible states. Cross-family donors test the boundary of abstraction. Matched-random donors preserve intervention scale and position without preserving causal identity. Local-API donors test whether operation-specific causal information can produce the target effect without the locality expected from a reusable interface. The complete donor–consumer matrix is reported so that role selectivity is visible rather than summarized by one diagonal average.

Natural use is tested through path-blocked mediation. For a target effect Δ_{total}, the mediated fraction is

$$\mathrm{MF} = \frac{\Delta_{\mathrm{total}} - \Delta_{\mathrm{blocked}}}{\Delta_{\mathrm{total}}}.$$

Candidate API paths are compared with null paths matched by family, layer distance, rank, activation norm, patchability, and intervention load. Necessity–restoration consistency tests whether the components whose ablation removes behavior are also those through which the correct interface code can restore it. Multi-consumer blocking tests whether one producer–consumer route supports several downstream operations rather than one output only.

Dimension and edge selectivity are evaluated with effect-size distributions, cross-seed replication, and BH-FDR correction. The graph is reported both at the pre-specified threshold and through threshold-free metrics such as AUPRC. Hub locations, sink locations, graph sparsity, edge frequency,

cross-seed Jaccard overlap, and threshold sensitivity are archived. A sparse graph is not accepted as mechanistic evidence unless its edge strength and mediated fraction exceed layer-, rank-, norm-, and patchability- matched null distributions.

Finally, the complete discovery and validation pipeline is applied blind to model organisms with known Shared, Local, Mixture, or Distributed routing. The primary falsification endpoint is the false-Shared discovery rate among non-Shared organisms. Overall class accuracy, correct-versus-runner-up MDL margin, posterior class mass, and overlap with the known communication graph are also reported. If non-Shared organisms are repeatedly classified as Shared, the coding contract, candidate alternatives, or fidelity thresholds must be revised before the real-model result can be interpreted. The output of the procedure is therefore either a selected architecture with its full coding, fidelity, intervention, graph, and falsification evidence, an explicitly ambiguous result, or a falsification result when no candidate satisfies the complete evidence contract.

X Extended Scope Conditions

Future causal signatures define a finite empirical quotient over the sampled operation bank. New operations can expose additional distinctions and refine existing equivalence classes; the reported interface is therefore indexed by the tested operation families, composition depths, output signatures, and distortion threshold.

The current operation bank is generated from controlled structured families. This construction provides known latent states, semantic-collision controls, and precise counterfactual futures, while open-domain dialogue, long-context retrieval, tool use, multimodal interaction, multilingual reasoning, and unconstrained scientific tasks introduce additional state and consumer structure beyond the present benchmark.

MDL rankings are comparative over the declared candidate family, prequential order, coding prior, parameter quantization, rank and gate penalties, and fidelity contract. Shared, Local, Mixture, Distributed, and the tested representation baselines define the present comparison; nonlinear, redundant, hierarchical, and dynamically routed interfaces remain distinct candidate families for future competition.

The empirical quotient is defined up to transformations that preserve future distributions and intervention effects. Rotations of a low-rank subspace can yield equivalent coordinates, so the result identifies a causal subspace and its organization rather than a canonical semantic basis for every dimension. Cross-seed alignment, principal-angle analysis, dimension selectivity, and re-encoding quantify this coordinate freedom.

Transplantation measures causal control under matched swaps. Nonlinear or redundant routes outside the candidate path can reduce the measured effect, whereas unusually patchable residual states can amplify it. Norm-, rank-, layer-distance-, and patchability-matched nulls characterize this intervention geometry, and locality metrics remain complemented by confidence-, sequence-, and path-level analyses.

Path mediation identifies a stable task-conditional causal backbone under finite interventions. Compensation through redundant routes can lower single-path mediation, while broad multi-path blocking can add nonspecific damage. Restoration, matched-load nulls, multi-consumer blocking, threshold sweeps, and threshold-free graph statistics define the measured route-level effect without treating the recovered graph as a complete circuit decomposition.

Known-routing model organisms calibrate the discovery rule under synthetic training. The $14/16$ recovery result includes one false-Shared prediction among twelve non-Shared organisms and path overlap 0.87 ± 0.06, providing an empirical specificity check for the tested regime. Large-scale pretrained mechanisms can contain more entangled objectives, nonlinear gates, and redundant pathways than these controlled organisms (Baker et al., 2026; Szablewski et al., 2026).

The two detailed language-model families provide cross-checkpoint replication, with a four-model directional sweep extending Sharedness Gain across additional backbones. Model scale, multilinguality, mixture-of-experts routing, recurrence, multimodality, instruction tuning, preference optimization, quantization, pruning, and adapters can alter layer location or interface geometry and define separate transfer regimes.

Capacity loading shows that the recovered channel is finite: success declines as simultaneously maintained causal classes increase and dual-task cost rises. Cross-family transplantation is also weaker than same-family transfer. Together these results support a hierarchy in which broadly reusable role structure coexists with family-specific interfaces and distributed computation.

Statistical precision is set by independent model and natural-family units rather than by the raw number of correlated prompt branches. Task-family bootstrap and matched seed analysis model this dependence; additional independently designed natural families and model families would tighten external-validity estimates.

Jensen–Shannon divergence is the primary fidelity metric. Its probability weighting can underemphasize rare decision-critical continuations, so answer-token, sequence-level, operation-tail, and worst-family analyses provide complementary summaries. Task-aware distortions constitute an alternative contract when fixed before architecture selection.

The study analyzes frozen pretrained checkpoints. Training-time emergence, data-driven pressure toward sharing, and reorganization during instruction or preference tuning are separate causal questions for longitudinal checkpoints and controlled training interventions. Within the tested models, layers, structured state families, operation bank, coding contract, and intervention protocol, Shared provides the shortest future-sufficient description and is reused by multiple measured downstream consumers.