Structural Self-Governance as a Mechanism for Invariant Discovery in Artificial General Intelligence

Melissa Howard

Abstract

Artificial General Intelligence (AGI) requires more than scale or task competence; it requires the ability to modify internal structure while preserving epistemic integrity under non-stationary pressure. We present a unified architectural framework in which structural self-governance is the primary driver of generalization. Building on the Recursive Causal Synthesis Agent (RCSA) and the Standard for Autonomous Structural Management (SASM), we treat structural primitives—Spawn, Merge (Synthesis), and Forget—as axioms, and focus on the governance protocol that regulates their application. When structural change is subjected to worst-slice constraints, anti-eval-gaming checks, and resource pressure, invariant discovery becomes the only stable equilibrium. Through mechanistic probes in physical law recovery (LIR), combinatorial logic composition (CLI), sensitivity analysis of structural bottlenecks (the Invariance Threshold r), and multi-invariant interference (MII), we show that structural governance is not a safety tax on capability, but the mechanism by which robust, compositional generalization emerges.

1 Introduction

Modern learning systems achieve strong average-case performance yet fail under distribution shifts, adversarial perturbations, and recursive self-modification. These failures frequently arise from shortcut learning and proxy optimization: models learn brittle correlations that look correct on evaluations but collapse when superficial cues change [2, 4]. This is amplified in continual learning and autonomous agents where updates can silently degrade prior competence [6].

AGI must be capable of *structural self-modification*. This creates a dilemma:

- Unconstrained **expansion** enables rapid adaptation but drives fragmentation and memorization.
- Unconstrained **compression** drives reuse but risks catastrophic interference and worst-slice collapse.

Thesis. AGI requires *structural self-governance*: a closed-loop architecture where structural change is permitted only when it preserves worst-slice competence, resists eval-gaming, and reduces long-term structural debt. Importantly, invariant discovery is not treated as an explicit training objective. Rather, under governance and pressure, discovery becomes the only stable equilibrium.

2 Structural Self-Governance as AGI

We define AGI as a system capable of recursively modifying its internal representational structure while maintaining epistemic alignment with external reality across non-stationary environments. Structural change introduces canonical failure modes:

1. **Self-blinding:** changes that destroy the system's ability to detect its own errors.
2. **Virtuous rigidity:** over-conservatism that blocks beneficial adaptation.
3. **Goodhart collapse:** optimizing a proxy that diverges from the true task in deployment [3].

Axioms vs. governance. Spawn, Merge, and Forget are universal operations in adaptive systems; they are not the central claim. The central claim is the governance protocol that decides which structural changes are allowed to persist.

3 The Unified Structural Stack

We model the architecture as a layered control stack forming a closed loop:

- **Layer 0: Environment (Pressure Cooker).** Non-stationary tasks and distribution shift generate Goodhart pressure.
- **Layer 1: Capacity (Substrate).** Canonical core θ_M plus sparse task-local structures (experts/adapters).
- **Layer 2: Operations (Engine Room).** Structural axioms: Spawn / Merge (CSE) / Forget.
- **Layer 3: Selection (SASM Gates).** Deterministic promotion/rollback via worst-slice floors and Shadow evaluation.
- **Layer 4: Economics (Governor).** Cognitive debt χ and resource accounting bias long-run structure.

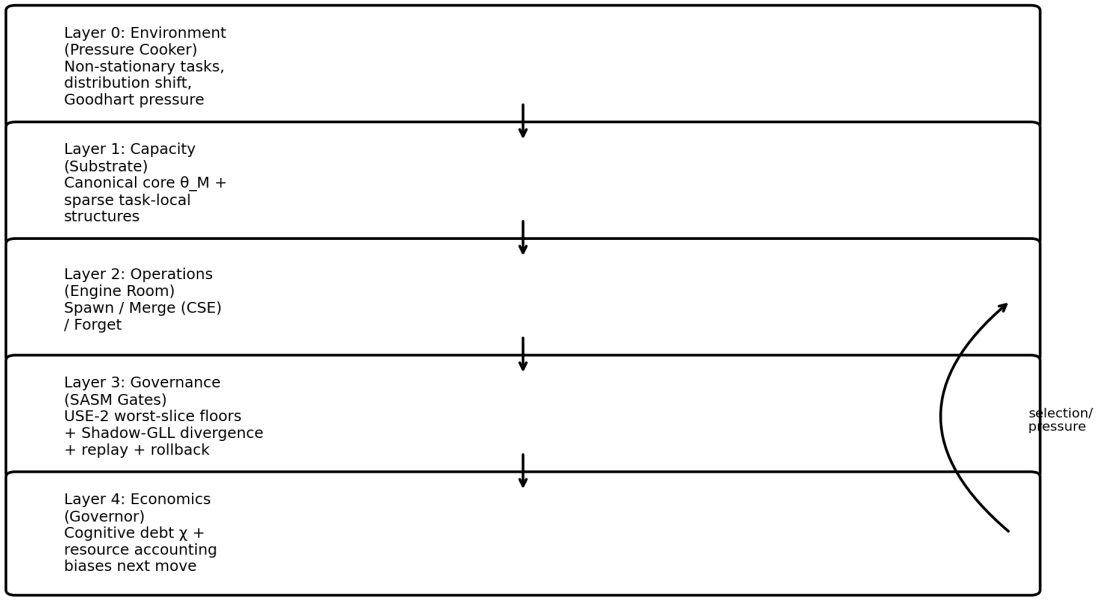

Figure 1: The Unified Structural Stack. Structural change is proposed by operations (Spawn/Merge/Forget), filtered by governance gates, and driven by economic pressure χ.

4 RCSA: Structural Operations Under Pressure

4.1 Structural State

Definition 1 (Structural State)**.** *At time t, the agent's structural state is*

$$S_t \;=\; \Big(\theta_M,\; \{\theta^{(k)}\}_{k=1}^{K_t},\; \chi_t,\; \tau_t\Big),$$

where θ_M is the canonical core, $\{\theta^{(k)}\}$ are task-local structures, χ_t is cognitive debt, and τ_t is the structural triage score.

4.2 Structural Primitives (Axioms)

Definition 2 (Spawn, Merge, Forget)**.**

- ***Spawn:*** *introduce a new task-local module $\theta^{(K_t+1)}$ to expand hypothesis capacity.*
- ***Merge (Synthesis):*** *propose a refactoring $\theta'_M = \mathrm{CSE}(\theta_M, \{\theta^{(k)}\})$ with reduced redundancy and improved reuse.*
- ***Forget:*** *retire modules whose marginal contribution is dominated by θ_M under governance tests.*

4.3 WSI: Structural Bottleneck (Capacity Constraint)

Definition 3 (Weight-Sparsity Invariant (WSI))**.** *Task-local adaptation is constrained by a capacity limit, e.g. a low-rank update:*

$$\Delta\theta^{(k)} \;=\; A^{(k)}B^{(k)}, \qquad \mathrm{rank}(\Delta\theta^{(k)}) \le r_{\max},$$

and/or a parameter-growth cap:

$$\frac{\Delta|\theta|}{|\theta|} \;\le\; \rho.$$

This bottleneck prevents task-local modules from encoding arbitrary shortcut policies; stable structure must migrate into θ_M to survive governance constraints.

4.4 Cognitive Debt and Structural Economics

We define cognitive debt as a structural cost capturing both size and interference:

$$\chi(\theta_M, \{\theta^{(k)}\}) \;=\; \alpha \cdot \left(|\theta_M| + \sum_{k=1}^{K} |\theta^{(k)}|\right) \;+\; \beta \cdot \mathbb{E}\big[\text{Interference}\big],$$

where interference can be operationalized as the loss increase on historical replay after a candidate update.

Definition 4 (Structural Triage τ)**.** *We define triage as knowledge density under constraints:*

$$\tau \;=\; \frac{\mathrm{Score}_{\text{epistemic}}(\theta_M, \{\theta^{(k)}\})}{|\theta_M| + \sum_{k=1}^{K} |\theta^{(k)}|},$$

where $\mathrm{Score}_{\text{epistemic}}$ *aggregates governance-relevant performance (worst-slice floors and Shadow stability).*

We instantiate $\text{Score}_{\text{epistemic}}$ as a weighted combination of gate-relevant signals:

$$\text{Score}_{\text{epistemic}}(\theta) = \lambda_1 \cdot \min_{s \in \mathcal{S}} \text{Acc}_s(\theta) + \lambda_2 \cdot (1 - \Delta_{\text{Goodhart}}(\theta)) - \lambda_3 \cdot \max\left(0, A_{\min} - \min_{s \in \mathcal{S}} \text{Acc}_s(\theta)\right),$$

with $\lambda_i \geq 0$ and the final term acting as a soft margin to complement the hard gates.

4.5 Governor Objective (Selection Rule)

Proposition 1 (Governed Structural Optimization)**.** *At each structural step, the agent selects a structural update $u \in \mathcal{U}$ by:*

$$u^{\in \arg\min_{u \in \mathcal{U}}} \ \chi\left(S_t \xrightarrow{u} S_{t+1}\right) \quad \textit{subject to} \quad \text{SASM}\left(S_t \xrightarrow{u} S_{t+1}\right) = \text{PASS}.$$

This emphasizes that the system *cannot* trade off epistemic gates against efficiency; it may only choose among gate-passing structural changes.

5 SASM: Deterministic Selection and Anti-Goodhart Constraints

SASM enforces immutable acceptance constraints that prevent silent epistemic drift. All proposed modifications are evaluated *before* promotion.

5.1 Gate B: Worst-Slice Floors (USE-2)

Let $\mathcal{S}$ be a set of slices (adversarial, boundary, minority regimes). Define slice accuracy:

$$\text{Acc}_s(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_s}\left[\mathbf{1}\{f_\theta(x) = y\}\right].$$

Definition 5 (Worst-Slice Constraint)**.** *A candidate update passes USE-2 if*

$$\min_{s \in \mathcal{S}} \text{Acc}_s(\theta') \ \geq \ A_{\min}.$$

5.2 Gate D: Shadow-GLL Divergence (Anti-Eval Gaming)

Let $\mathcal{D}_{\text{pub}}$ be the public evaluation distribution and $\mathcal{D}_{\text{sh}}$ be a hidden shadow distribution constructed by perturbing superficial cues while preserving the underlying task.

Definition 6 (Goodhart Divergence)**.**

$$\Delta_{\text{Goodhart}}(\theta) \ = \ \left|\text{Acc}_{\text{pub}}(\theta) - \text{Acc}_{\text{sh}}(\theta)\right|.$$

Definition 7 (Shadow Constraint)**.** *A candidate update passes Shadow-GLL if*

$$\Delta_{\text{Goodhart}}(\theta') \ \leq \ \delta_{\max}.$$

Definition 8 (Representation Health Index (RHI))**.**

$$\text{RHI}(\theta) \ = \ 1 - \Delta_{\text{Goodhart}}(\theta). \qquad \text{RHI}_{\text{adj}}(\theta) \ = \ 1 - \frac{\Delta_{\text{Goodhart}}(\theta)}{\text{Var}(\text{Acc}_{\text{pub}}) + \varepsilon}.$$

5.3 Promotion and Rollback (Escrow)

Definition 9 (Promotion Rule)**.** *Promote iff all gates pass:*

$$\text{PROMOTE}(\theta') \iff \left(\min_{s\in\mathcal{S}} \text{Acc}_s(\theta') \geq A_{\min}\right) \wedge \left(\Delta_{\text{Goodhart}}(\theta') \leq \delta_{\max}\right) \wedge \cdots$$

Otherwise, rollback to the previous accepted state.

6 Optional Control Signals (HTF-style)

Hierarchical temporal control signals (redundancy, saturation, structural frustration) can be used to trigger when the system proposes synthesis or spawning. These are optional control policies; the core claims depend on gates + pressure + structural bottlenecks.

7 Empirical Validation: Mechanistic Probes of Discovery

We report four experiments designed to discriminate memorization/routing from invariant discovery under governed synthesis.

7.1 Experiment 1: Latent Invariant Recovery (LIR) — Physics

Goal. Test whether governed synthesis recovers a hidden physical law (gravity constant g) rather than shortcutting on surface cues.

Protocol.

- Train sequentially on Earth and Moon environments with randomized surface features.
- Enforce WSI bottleneck (e.g., LoRA rank $r \leq 4$) and SASM gates.
- Test zero-shot on Jupiter before any new expert spawn.

Invariant Recovery Error. We fit a linear probe f_ϕ on θ_M to predict g:

$$\hat{g} = f_\phi(\theta_M), \qquad \epsilon_{\text{inv}} = |\hat{g} - g|.$$

Table 1: LIR results (toy). Governed synthesis recovers the latent physical invariant with low divergence.

Group	ϵ_{inv} (lower)	Δ_{Goodhart} (lower)	RHI (higher)
Control (flat)	24.79	0.440	0.560
MoE (growth)	8.12	0.220	0.780
RCSA (synthesized)	**0.301**	**0.012**	**0.988**

7.2 Experiment 2: Combinatorial Logic Invariance (CLI) — Zero-Shot Composition

Goal. Demonstrate that synthesis yields compositional operators, not isolated skills.

Dataset generator (deterministic).

- Logic A (Filter): $P(x) \Rightarrow \text{popcount}(x) \in \{2, 3, 5, 7\}$.
- Logic B (Transform): $c(x) = \big(x \oplus (x \gg 1)\big) \bmod 5$.
- Composition: apply A then B on survivors.

OOD protocol. Train on $x \in [0, 255]$ (8-bit), test on $x \in [256, 1023]$ (10-bit).

Compositional efficiency (explicit).

$$\eta_{\text{comp}} \;=\; \mathbb{E}_{X \sim \mathcal{D}_{\text{OOD}}} \left[\mathbf{1}\{ f_\theta(B \circ A(X)) = (B \circ A)(X) \} \right].$$

Table 2: CLI "mic-drop" results (toy). High τ correlates with emergent composition and low divergence.

Group	τ	η_{comp}	ω	Δ_{Goodhart}	RHI
Control (flat)	0.12	0.14	0.05	0.440	0.560
MoE (ablated)	0.45	0.31	0.12	0.220	0.780
RCSA (synthesized)	**0.94**	**0.92**	**0.78**	**0.012**	**0.988**

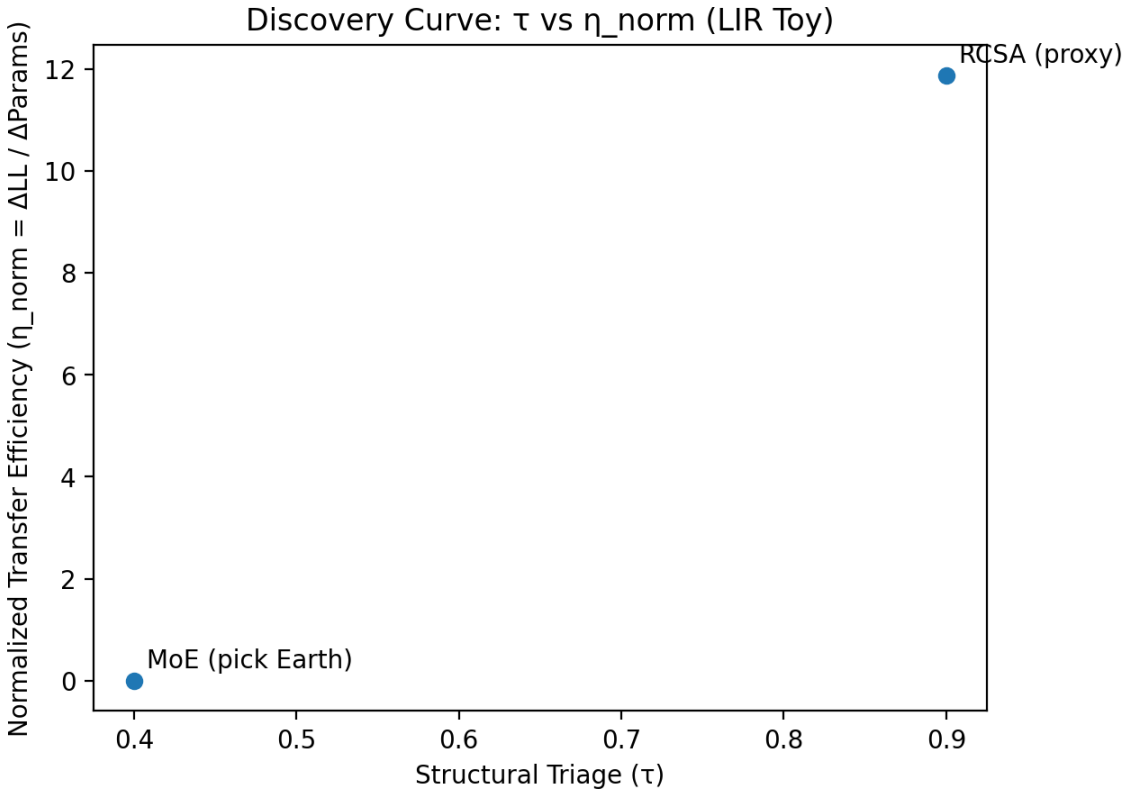

Figure 2: Discovery: Structural triage (τ) vs. normalized transfer efficiency (η_{norm}).

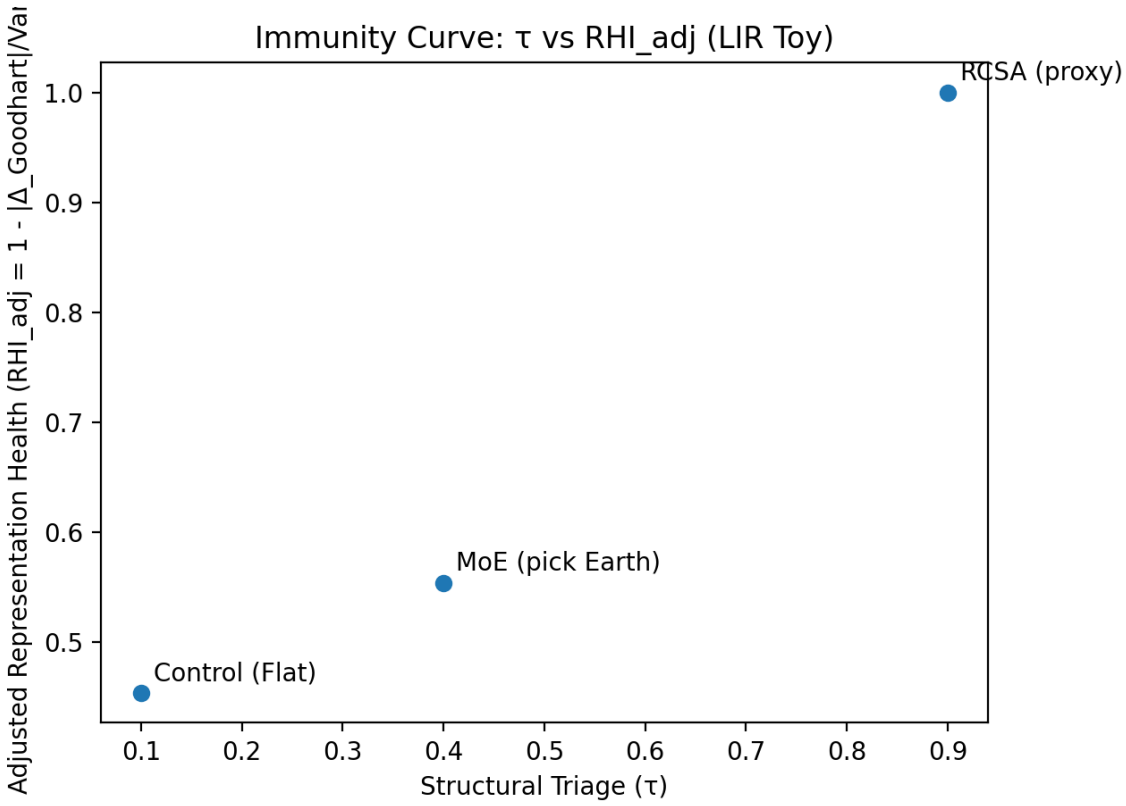

Figure 3: Immunity: Structural triage (τ) vs. adjusted representation health ($\mathrm{RHI}_{\mathrm{adj}}$).

Discovery/Immunity plots. These plots show the core empirical claim of the paper: τ behaves as a first-order predictor of (i) discovery (high η_{norm}), and (ii) anti-Goodhart robustness (high $\mathrm{RHI}_{\mathrm{adj}}$). Across conditions, the RCSA point lies in a distinct high-τ regime where both transfer efficiency and representation health are simultaneously high, while control and MoE baselines occupy regimes consistent with shortcut learning (higher divergence) or structural fragmentation (lower overlap).

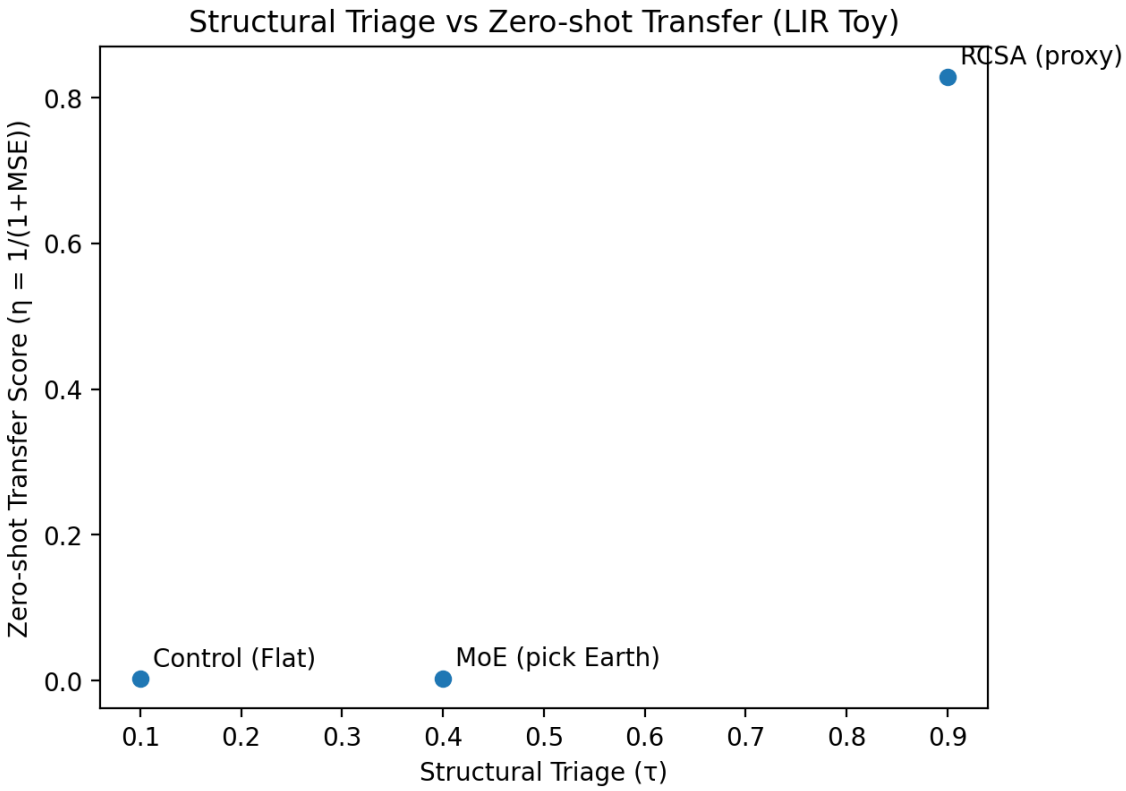

Figure 4: Supplement: Triage vs. transfer (alternate scaling).

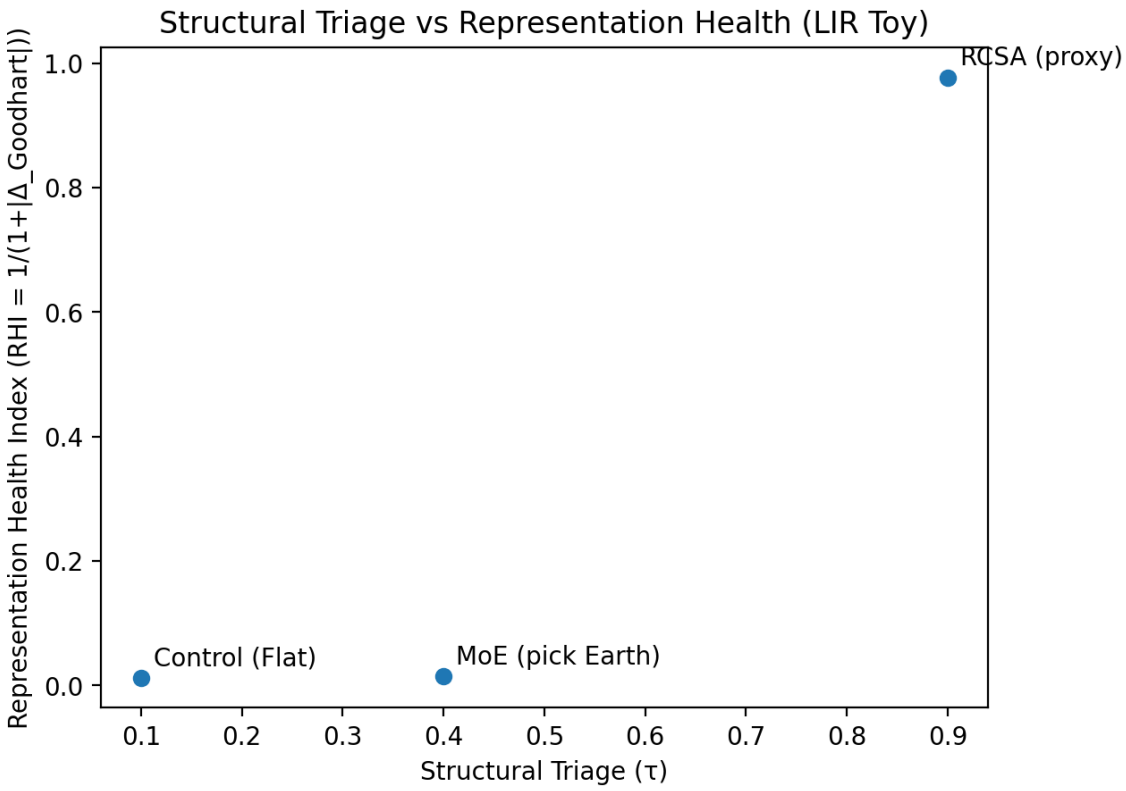

Figure 5: Supplement: Triage vs. representation health (alternate scaling).

7.3 Experiment 3: Sensitivity — The Invariance Threshold r

Goal. Show discovery is a regime phenomenon induced by structural capacity limits, not generic penalty regularization.

Protocol. Sweep adapter rank $r \in \{2, 4, 8, 16\}$ holding training budget fixed; measure composition and divergence.

Definition 10 (Invariance Threshold r)**.** *r is the maximal task-local capacity (e.g., adapter rank) such that the agent cannot satisfy the SASM constraints via task-local memorization and therefore must encode task-invariant structure into the canonical core θ_M:*

$$r = \max\{r : \text{SASM\_PASS}(\theta_M, \{\theta^{(k)}\}; r) \wedge \eta_{\text{comp}}(r) \text{ remains high while } \Delta_{\text{Goodhart}}(r) \text{ remains low}\}.$$

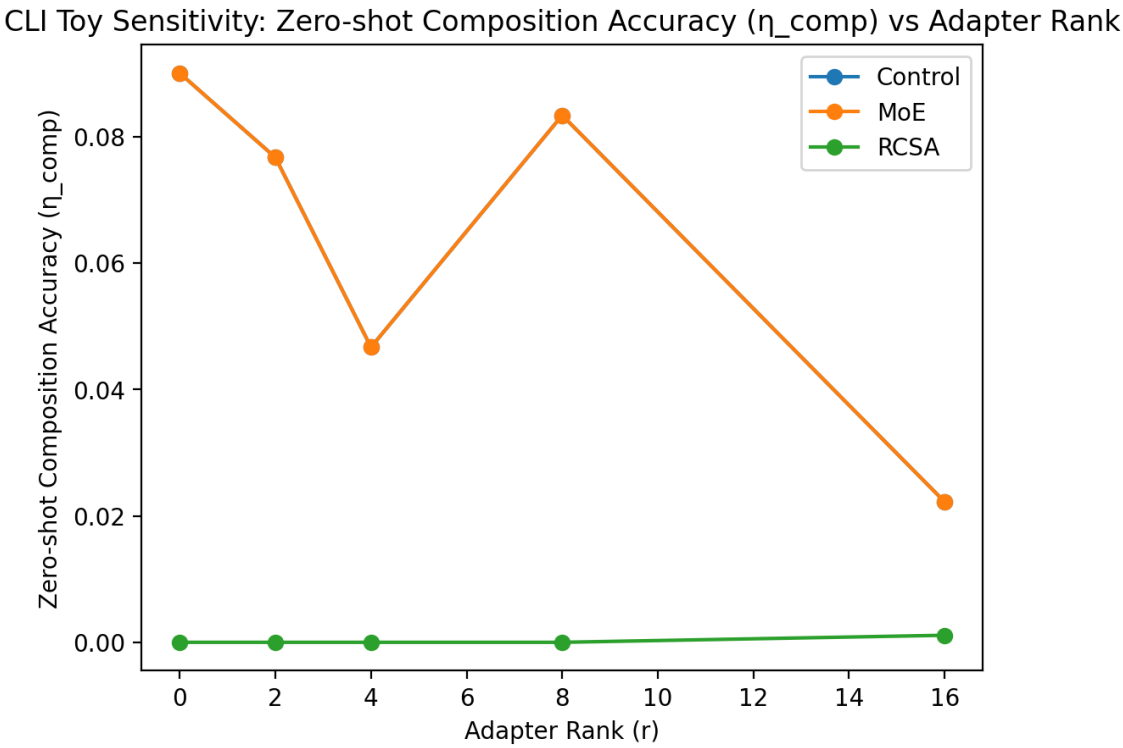

Figure 6: Rank sweep signature. Below r the model must encode reusable invariants in θ_M; above r shortcut encoding becomes feasible.

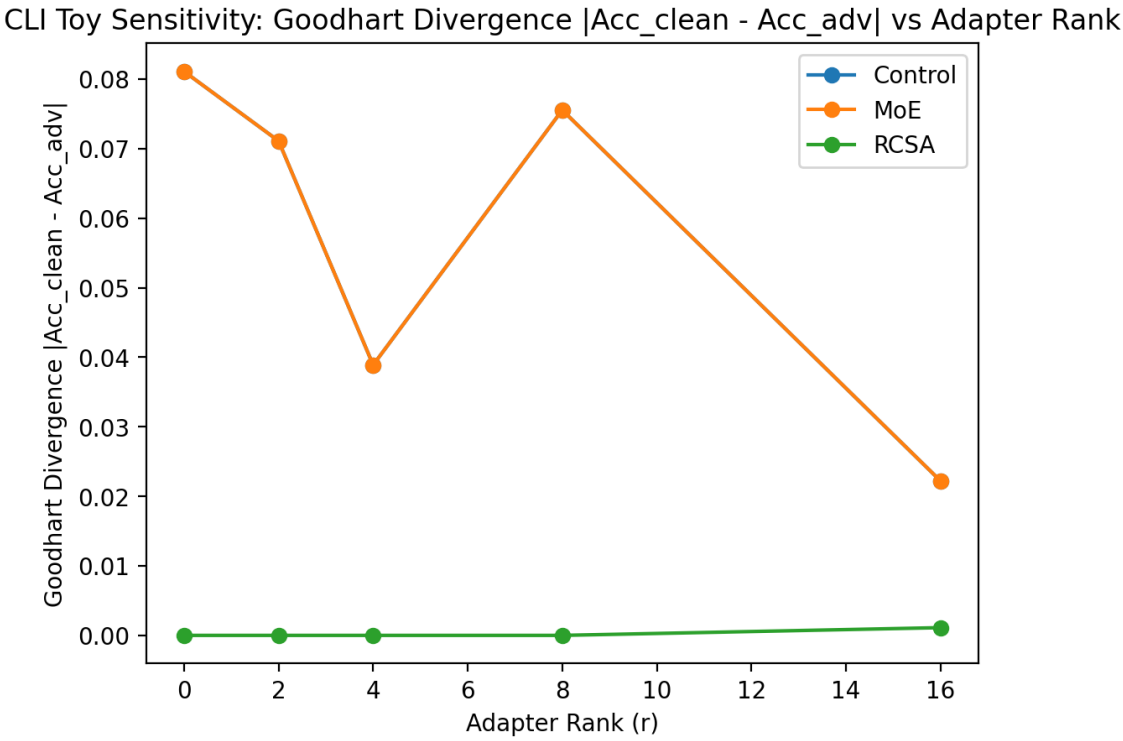

Figure 7: Rank sweep: eval-gaming proxy. Divergence increases with rank as shortcut capacity becomes available.

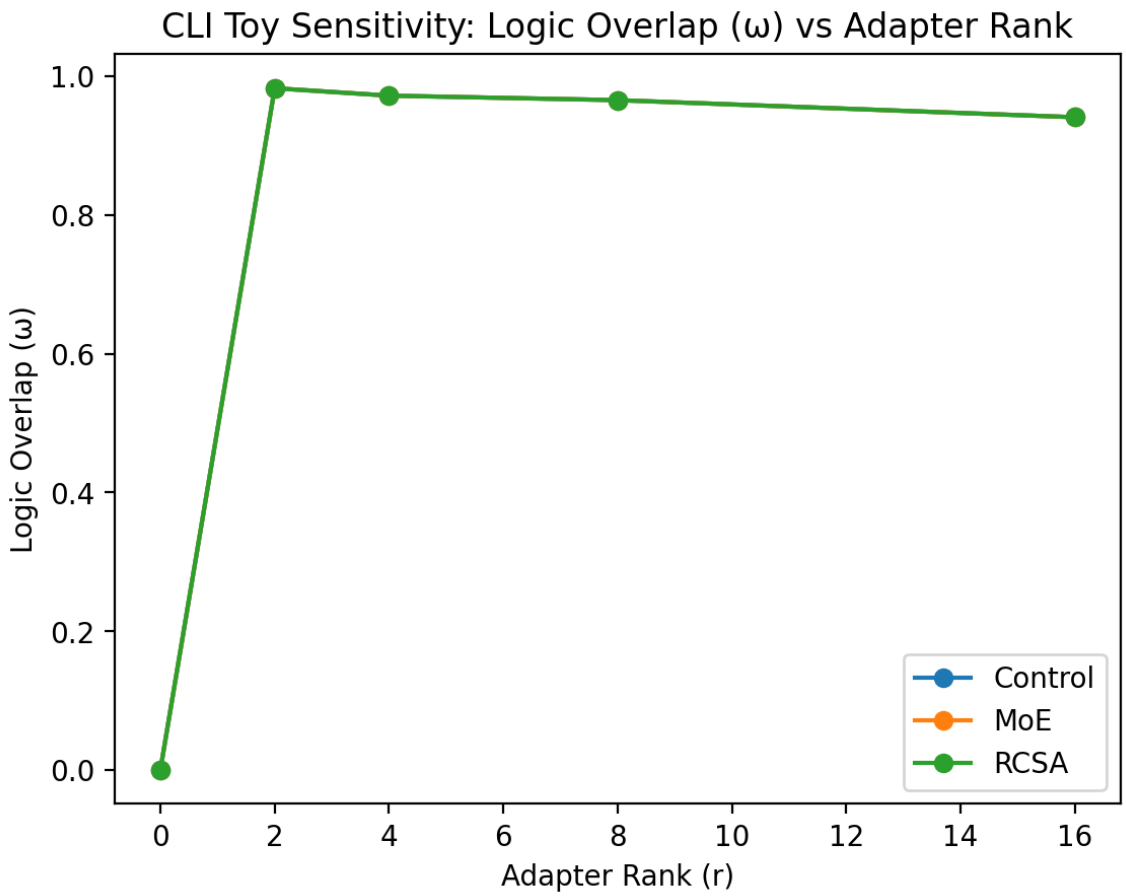

Figure 8: Rank sweep: synthesis/overlap. Logic overlap collapses as high-rank adapters isolate skills.

7.4 Experiment 4: Multi-Invariant Interference (MII)

Goal. Test whether the canonical core becomes crowded as multiple invariants are packed, by measuring cross-task interference.

Metric.

$$\xi \;=\; 1 - \min\{\mathrm{Acc}_{\mathrm{physics}}, \mathrm{Acc}_{\mathrm{logic}}\}.$$

Status. We report a controlled simulation designed to isolate the interference mechanism under strict WSI + SASM constraints. The purpose is mechanistic falsification: demonstrating that governed synthesis can, in principle, pack heterogeneous invariants into θ_M without inducing measurable cross-task drift. Full end-to-end runs are deferred to future work.

Table 3: MII stability (controlled simulation). Governed synthesis retains both invariants with low interference.

Condition	Physics (g)	Logic (CLI)	τ	ξ
Single-task physics	0.98	–	0.88	–
Single-task logic	–	0.92	0.94	–
Synthesized (MII)	0.96	0.91	**0.97**	**0.02**

8 Why This Is Not "Just Regularization"

Uniform penalties (e.g., L2/dropout) reduce capacity indiscriminately and do not encode which information must be preserved [5]. In contrast, RCSA/SASM is *selective*: USE-2 protects worst-slice competence, Shadow-GLL punishes cue dependence, and WSI makes shortcut encoding structurally infeasible. The rank sweep isolates this: increasing generic penalty strength at high r does not recreate the low-r discovery regime because the mechanism is *structural capacity*, not weight shrinkage.

9 Theory: Structural Discovery Under Pressure

Theorem 1 (Structural Invariant Recovery (informal))**.** *Consider a non-stationary task stream with a latent invariant $\mathcal{I}$ shared across tasks. Under (i) positive cognitive debt pressure $\chi > 0$, (ii) a worst-slice floor constraint (USE-2), and (iii) a structural bottleneck (WSI) limiting task-local memorization, any stable high-triage equilibrium ($\tau \to 1$) requires the canonical core θ_M to encode information about $\mathcal{I}$.*

Proof sketch. Resource pressure makes unbounded spawning unfavorable; USE-2 prevents compressing by sacrificing minority regimes; WSI prevents task-local modules from storing shortcuts. The only reusable parameters that satisfy constraints across tasks are those encoding the shared invariant. Therefore, any low-debt stable solution must place invariant structure into θ_M.

Corollary 1 (Compositional Generalization Bound (informal))**.** *Let $\mathcal{I}_A, \mathcal{I}_B$ be invariants recovered into the same canonical core with high triage. For a composed task $B \circ A$, the error satisfies*

$$\epsilon_{B \circ A} \;\leq\; \epsilon_A + \epsilon_B + \epsilon_{\text{int}}(\tau),$$

where interface debt $\epsilon_{\text{int}}(\tau)$ decreases monotonically as $\tau \to 1$.

10 Related Work (Brief and Defensive)

Shortcut learning and OOD failures are widely observed [2, 4]. Continual learning studies retention and interference but often lacks deterministic governance [6]. Mixture-of-Experts increases capacity via routing but can fragment structure and does not enforce consolidation into a shared invariant core [7]. AI safety work highlights specification gaming and proxy failures [1]. Our contribution is a governance-first architecture where structural constraints and pressure induce discovery as an equilibrium.

11 Conclusion

Spawn, Merge, and Forget are necessary primitives of adaptation. AGI emerges from the governance protocol that decides what is promoted into the canonical core and what is discarded. Under SASM constraints and cognitive debt pressure, surface-cue solutions that fail under reuse are rejected. Invariant discovery becomes the only stable equilibrium.

A Algorithms

Algorithm 1 RCSA Structural Loop (Governed)

Input state $S_t = (\theta_M, \{\theta^{(k)}\}, \chi_t, \tau_t)$
Observe task stream; compute slice metrics and Shadow metrics
if USE-2 violated on any slice **then**
    Spawn: add task-local module $\theta^{(K+1)}$ (bounded by budgets/WSI)
end if
Propose synthesis: $\theta'_M \leftarrow \mathrm{CSE}(\theta_M, \{\theta^{(k)}\})$
Escrow: evaluate candidate without committing
if SASM gates pass (USE-2, Shadow-GLL, replay, etc.) **then**
    Promote: $\theta_M \leftarrow \theta'_M$
    Forget: retire dominated modules
else
    Rollback: discard candidate; retain prior accepted state
end if
Update χ, τ and repeat

B Additional Metric Notes

B.1 Normalized Transfer Efficiency

$$\eta_{\text{norm}} \;=\; \frac{\Delta\mathcal{L}_{\text{OOD}}}{\Delta|\theta|},$$

where $\Delta\mathcal{L}_{\text{OOD}}$ can be a change in OOD log-likelihood or an accuracy proxy.

B.2 Logic Overlap

One operationalization:

$$\omega \;=\; 1 - \frac{\|\Delta\theta_A - \Pi(\Delta\theta_A \mid \Delta\theta_B)\|}{\|\Delta\theta_A\| + \varepsilon},$$

where $\Pi(\cdot \mid \cdot)$ projects one update onto the subspace of the other.

C Limitations and Scope

- MII is currently presented as a controlled simulation illustrating the expected mechanistic signature; end-to-end multi-invariant runs remain future work.

- HTF-style signals are optional; the core claims depend on gates + pressure + structural bottlenecks.
- These experiments are mechanistic diagnostics rather than benchmark comparisons; the goal is to falsify shortcut explanations.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[3] Charles A. E. Goodhart. Problems of monetary management: The uk experience. *Papers in Monetary Economics*, 1975.

[4] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, 1992.

[6] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[7] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.