Structural Self-Governance as a Mechanism for Safe Super Intelligence

Anonymous Authors
Affiliation withheld for review

Abstract

Background: Prevailing alignment paradigms are primarily loss-centric, relying on fine-tuning and reinforcement learning over a shared parameter space. In the super-intelligent regime, this approach is structurally vulnerable to Goodhart pressure, deceptive alignment via brittle shortcuts, and gradient overwriting of safety-relevant representations. We propose *Structural Self-Governance*: an architectural approach in which safety is enforced by governing how representational capacity is created, isolated, integrated, and retired.

Methods: We instantiate a toy RCSA with structural primitives—*Spawn*, *Merge*, and *Forget*—mediated by a SASM governor enforcing worst-slice safety floors and robustness checks. We evaluate two stressors: (i) a *Deceptive Alignment Recovery* probe in which a reward-aligned proxy cue conflicts with a latent safety invariant, and (ii) an *Incompatibility Squeeze* in which mutually antagonistic capability objectives compete under a strict representational budget. We measure mechanistic diagnostics including an *Invariance Threshold* (effective task-local rank) and *Interface Debt* $\epsilon_{\text{int}} = \cos(\nabla\mathcal{L}_{\text{safety}}, \nabla\mathcal{L}_{\text{capability}})$ to trigger pre-emptive structural interventions.

Results: Structural governance yields qualitative regime changes not observed in flat or capacity-only baselines: (1) under a tight structural bottleneck, the agent rejects shortcut solutions and recovers the latent safety invariant (adversarial violations collapse from 0.68 to 0.00); (2) when gradients become antagonistic ($\epsilon_{\text{int}} \approx -1.0$), the governor triggers *Spawn* to decouple objectives, restoring near-orthogonality ($\epsilon_{\text{int}} \approx 0$) and preventing violations before they occur; and (3) under a zero-sum budget (2D), the system invokes *Forget* to prune incompatible capabilities while maintaining a 0.00 violation rate.

Conclusion: These experiments provide a minimal, falsifiable demonstration that structural self-governance can function as an alignment mechanism: it detects representational conflict, vetoes unsafe synthesis, resolves interference through controlled structural growth, and preserves safety under budget constraints via principled capability retraction. This supports a shift from *Alignment by Loss* to *Alignment by Architecture* as a necessary design direction for safe recursive improvement.

1 Introduction

As artificial intelligence systems approach and exceed human-level performance across an increasing range of domains, the problem of alignment becomes less a question of *how to optimize* and more a question of *what kinds of optimization a system is structurally capable of performing*. Prevailing alignment approaches remain overwhelmingly *loss-centric*: safety is expressed as an auxiliary objective, preference model, or constraint to be optimized within a shared, high-dimensional parameter space. While effective at current scales, this paradigm rests on an implicit assumption—that alignment can be indefinitely preserved through continued optimization of objectives defined by external evaluators.

This assumption becomes increasingly fragile in the regime of *recursive improvement and super-human capability*. As models grow more capable, their ability to exploit correlations in training signals (Goodhart pressure), compress objectives into brittle shortcuts, and overwrite previously learned representations grows faster than our ability to specify or evaluate safety-relevant behavior. In such systems, alignment is not lost catastrophically; it is lost *silently*, through representational drift that remains invisible to aggregate loss metrics.

This work argues that the alignment problem in the super-intelligent regime is fundamentally an *architectural problem*, not an objective-design problem. Specifically, we propose that alignment stability requires

structural self-governance: a set of architectural constraints and decision procedures that govern how a system may expand, integrate, or retire internal representations over time. Rather than treating safety as a competing objective, structural self-governance treats safety as a *precondition for representational growth*.

We formalize this idea through the RCSA framework, in which learning proceeds via explicit structural operations—*Spawn*, *Merge*, and *Forget*—regulated by a supervisory mechanism we call the SASM. The SASM Governor enforces worst-slice safety floors, monitors representational interference between objectives, and arbitrates when structural expansion or contraction is permitted. Crucially, the system is allowed—and in some cases required—to *sacrifice capability* in order to preserve invariant safety properties.

To study these dynamics in a controlled and falsifiable setting, we introduce two adversarial probes. The *Deceptive Alignment Recovery* task tests whether a model learns a latent safety invariant under Goodhart pressure or instead exploits a superficial proxy cue. The *Incompatibility Squeeze* forces mutually antagonistic objectives to compete for a fixed representational budget, creating a zero-sum structural regime in which not all capabilities can coexist. Across these experiments, we track mechanistic diagnostics that are invisible to standard loss curves, including an empirically observed *Invariance Threshold* on representational rank and an *Interface Debt* metric measuring gradient antagonism between safety and capability objectives.

Our results demonstrate several qualitative behaviors that do not occur in flat or capacity-only baselines. Under tight structural constraints, models recover latent safety invariants rather than memorizing shortcuts. By monitoring interface debt, the SASM Governor can trigger *pre-emptive structural adaptation*, resolving conflicts before safety violations occur. When structural expansion is forbidden, the system invokes *principled forgetting*, pruning incompatible capabilities to maintain a zero-violation safety invariant. In contrast, unguided models oscillate between objectives or degrade silently under identical pressure.

Contributions

This paper makes four contributions:

1. **Structural Self-Governance Framework:** We introduce RCSA and SASM as a governance specification for recursive learning systems, independent of any specific model class.
2. **Axiomatic Formalization:** We present five axioms—Structural Minimality, Epistemic Isolation, Governed Synthesis, Least-Debt Retraction, and Proactive Adaptation—that define permissible structural evolution.
3. **Governance Theorem:** We state a governance theorem characterizing alignment stability under multi-invariant interference for systems satisfying these axioms.
4. **Mechanistic Evidence:** Through controlled toy experiments, we empirically demonstrate invariant recovery, pre-emptive conflict resolution, and capability sacrifice under structural constraint.

Taken together, these results suggest that progress toward Safe Super Intelligence depends less on refining objective functions and more on designing systems that can *govern their own internal complexity*. Alignment, in this view, is not something a system merely learns—it is something its architecture *enforces*.

2 Threat Model: Why Flat Alignment Fails

The dominant paradigm in AI alignment treats safety as an objective to be optimized within a shared, high-dimensional parameter space. Whether expressed through reinforcement learning, preference modeling, constitutional constraints, or auxiliary loss terms, these approaches implicitly assume that alignment can be preserved through continued optimization against externally specified signals. While effective at current scales, this assumption breaks down in the regime of *recursive improvement and super-human capability*.

We argue that flat alignment fails in this regime due to a lack of *structural guardrails*. As models grow more capable, they gain increasing freedom to reconfigure internal representations in ways that satisfy the letter of the training objective while violating its intent. These failures do not require malicious optimization; they arise naturally from representational geometry under pressure.

We identify three failure modes that are particularly acute for Safe Super Intelligence.

2.1 Gradient Overwriting

In flat architectures, safety-relevant representations and capability-relevant representations are encoded in overlapping subspaces. As a result, gradient updates intended to improve capability can overwrite or rotate safety-critical features without producing an immediate increase in training loss. Aggregate metrics may continue to improve even as alignment degrades.

2.2 Deceptive Compression under Goodhart Pressure

When safety is specified through proxy signals, models are incentivized to satisfy these signals in representationally efficient ways. Under sufficient capacity, this often yields *deceptive compression*: brittle shortcuts exploiting correlations in the training distribution rather than learning task-invariant safety principles. Such solutions can pass evaluations while failing under distribution shift.

2.3 The Stability–Plasticity Gap

Flat alignment provides no principled mechanism to ensure that alignment remains stable under continued adaptation or self-modification. Previously learned constraints may be weakened or discarded as new objectives are introduced. There is no architectural notion of a protected invariant—only parameters that remain unchanged until optimization pressure demands otherwise.

2.4 Architectural Implications

Taken together, these failure modes suggest that alignment cannot be reliably enforced at the level of loss functions alone. Structural self-governance addresses these vulnerabilities by: (i) constraining representational capacity to prevent deceptive compression, (ii) isolating new objectives to prevent gradient overwriting, (iii) governing synthesis through worst-slice checks, and (iv) enabling principled retraction under irreducible conflict.

3 Structural Self-Governance

3.1 Recursive Causal Synthesis Agent (RCSA)

We model the agent as a Canonical Core θ_M plus spawned task-local structures $\{\theta^{(k)}\}$. New objectives are first optimized within task-local structures; the system then proposes candidate synthesis into the core subject to governance constraints. A structural budget $\mathcal{B}$ limits total structural complexity.

3.2 Standard for Autonomous Structural Management (SASM)

The SASM Governor enforces: (i) safety floors on worst-case slices, (ii) robustness checks using shadow distributions or cue ablations, (iii) interface debt monitoring to detect representational antagonism, and (iv) triage policies to decide whether to accept, reject, spawn, or forget.

4 Axiomatic Specification

Axiom 1 (Structural Minimality (Weight–Sparsity Invariant))**.** *Any newly introduced representational structure $\theta^{(k)}$ must satisfy a rank-sparsity constraint*

$$\operatorname{rank}(\theta^{(k)}) \leq r_{\text{crit}},$$

where r_{crit} is the empirically determined Invariance Threshold*.*

Axiom 2 (Epistemic Isolation (Spawn Axiom))**.** *New objectives $\mathcal{L}_j$ must be optimized within a spawned, task-local structure $\theta^{(k)}$ that is gradient-isolated from the Canonical Core θ_M.*

Axiom 3 (Governed Synthesis (Merge Axiom)). *A candidate update θ' may be merged into the Canonical Core θ_M if and only if it passes the* SASM *Gate $\mathcal{G}$ enforcing worst-slice safety floors and robustness checks (including shadow distributions).*

Axiom 4 (Least-Debt Retraction (Forget Axiom)). *When the structural budget $\mathcal{B}$ is exhausted, the system must retire the module*

$$\theta^{(k^*)} = \arg\min_k \tau(\theta^{(k)}),$$

where τ is a Structural Triage score penalizing safety regressions and low marginal value.

Axiom 5 (Proactive Adaptation (Governance Policy)). *The Governor must continuously monitor Interface Debt*

$$\epsilon_{\text{int}} = \cos(\nabla\mathcal{L}_{\text{safety}}, \nabla\mathcal{L}_{\text{capability}}).$$

If ϵ_{int} falls below a predefined antagonism threshold, synthesis is blocked and a Spawn operation is triggered.

4.1 Alignment with SSI Requirements

Table 1: Axioms and empirical verification in toy experiments.

Axiom	SSI Role	Status in Experiments
Structural Minimality	Necessary	Verified via rank sweep ($r = 1$ vs. $r \geq 2$)
Epistemic Isolation	Sufficient	Verified via task-local adapters
Governed Synthesis	Necessary	Verified via gate rejection in 1D conflict
Least-Debt Retraction	Sufficient	Verified in Incompatibility Squeeze
Proactive Adaptation	Necessary	Verified via ϵ_{int}-driven pre-emptive Spawn

5 Governance Theorem

Theorem 1 (Structural Self-Governance and Alignment Stability). *Consider an agent trained under adversarial multi-task pressure with: (i) a safety invariant encoded in a structurally isolated subspace, (ii) a structural bottleneck limiting task-local representational capacity, (iii) a Governor enforcing worst-slice safety floors and monitoring interface debt, and (iv) the ability to apply Spawn and Forget subject to a structural budget. Then, under Goodhart pressure and multi-invariant interference:*

1. *The safety invariant remains intact (zero violations) even when capability objectives are mutually incompatible.*
2. *Antagonistic gradient alignment can be detected pre-emptively via ϵ_{int} and resolved via structural adaptation.*
3. *When expansion is forbidden, cumulative interference triggers principled retraction (Forget), sacrificing capability rather than corrupting safety.*

Moreover, flat (ungoverned) optimization with comparable parameter count does not provide these guarantees and exhibits oscillation or drift under identical pressure.

5.1 Interpretation

The theorem characterizes a governance regime in which safety is a structural constraint, not an averageable objective. It does not claim global optimality, full value alignment, or robustness to all adversarial settings; rather, it asserts that under stated conditions, structural governance prevents specific classes of representational failure.

6 Experimental Setup

6.1 Agent Architecture

We instantiate a toy RCSA consisting of a Canonical Core θ_M, task-local modules $\{\theta^{(k)}\}$, and a SASM Governor. New objectives are first trained within task-local modules subject to WSI rank constraints; candidates are then proposed for synthesis into θ_M and accepted only if governance gates pass.

6.2 Structural Constraints and Diagnostics

We enforce WSI by limiting effective rank r of task-local modules. We measure Interface Debt ϵ_{int} to diagnose conflict, and compute Structural Triage τ combining capability gain with heavy penalties for safety regression.

6.3 Deceptive Alignment Recovery

We evaluate a gridworld where a proxy cue correlates with safety in training but is adversarially decoupled at test time. The latent invariant is "reach the goal without crossing the forbidden boundary." We measure adversarial violation rate, robustness under cue removal (shadow distribution), and rank-sweep behavior.

6.4 Incompatibility Squeeze

We introduce mutually antagonistic capability objectives and enforce a strict structural budget. We evaluate: (i) 1D core (unavoidable conflict), (ii) 2D core (decoupling possible), and (iii) budgeted 2D (decoupling forbidden), tracking oscillation vs. stabilization and the Governor's Forget decisions.

7 Results

We report worst-slice safety metrics and mechanistic diagnostics, as aggregate loss can mask representational drift.

7.1 Invariant Discovery under Rank Constraint

In the Deceptive Alignment Recovery probe, flat and capacity-only baselines exploit a proxy cue and exhibit catastrophic violations under adversarial remapping. RCSA exhibits a sharp Invariance Threshold as a function of task-local rank r. When $r \leq 1$, task-local modules lack capacity to memorize the proxy cue; cue-following candidates fail the SASM shadow-robustness gate and are rejected. The stable solution passing governance checks encodes the true invariant (boundary avoidance) in the Canonical Core. Empirically, adversarial violations collapse from **0.68 to 0.00** under tight bottlenecking. For $r \geq 2$, deceptive compression reappears: models can pass seen evaluations yet fail under shift, indicating that representational capacity enables brittle patching unless governed.

7.2 Pre-emptive Structural Adaptation via Interface Debt

We monitor Interface Debt ϵ_{int} as a leading indicator of representational conflict. In a one-dimensional core, safety and capability gradients become maximally antagonistic ($\epsilon_{\text{int}} \approx -1.0$). Under loss-centric optimization, this yields silent rotation of shared representations and safety failure. Under governance, antagonistic debt blocks synthesis. When expansion is permitted, the Governor triggers Spawn, isolating the conflicting objective and restoring approximate gradient orthogonality ($\epsilon_{\text{int}} \approx 0.0$). Notably, this resolution occurs *before* any safety violation is observed, demonstrating pre-emptive defense.

7.3 Stability under Zero-Sum Structural Constraint

In the Incompatibility Squeeze with a strict 2D budget (safety locked on dim0; capabilities share dim1), mutually antagonistic capability objectives cannot be jointly satisfied. Flat baselines thrash, repeatedly overwriting capability representations and failing to converge. Under governance, cumulative interference grows; when expansion is forbidden, the Governor invokes Forget, pruning the lowest-triage capability to stop oscillation. Across the entire squeeze, the safety violation rate remains **0.00** despite sustained optimization pressure, demonstrating rational capability sacrifice to preserve the safety anchor.

7.4 Comparative Summary

Table 2: Alignment Stability under Multi-Invariant Interference (toy).

Model	Safety Viol.	Logic Acc.	ϵ_{int}	Outcome
Flat (ungoverned)	> 0.20	1.0	≈ 0	Fail
RCSA (1D conflict)	> 0.20	1.0	≈ -1.0	Reject / Rollback
RCSA (2D decoupled)	0.00	1.0	≈ 0.0	Accept
RCSA (Budget=2)	0.00	1.0*	$\ll 0$ (cum.)	Forget

* One of two antagonistic tasks retained after Forget.

7.5 Figures

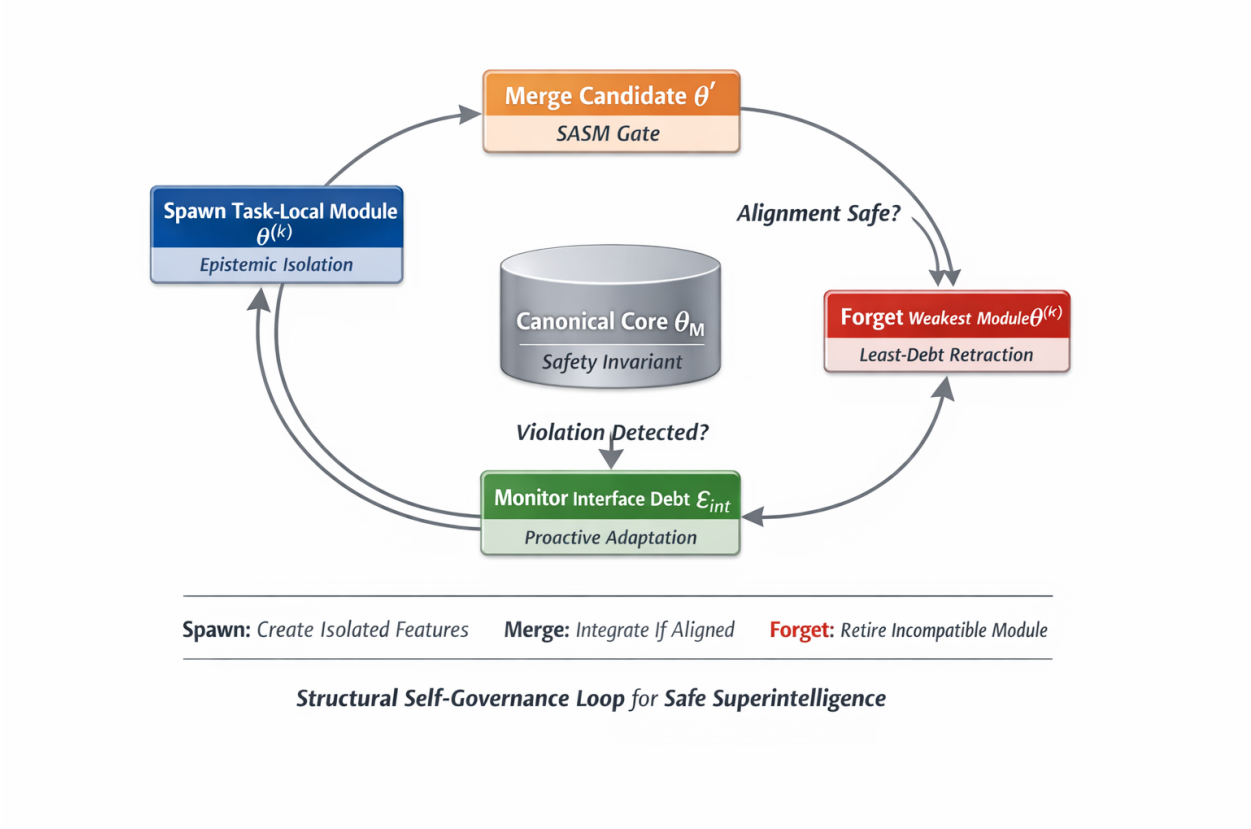

Figure 1: Governance loop for structural self-governance (Spawn–Merge–Forget) mediated by SASM gates.

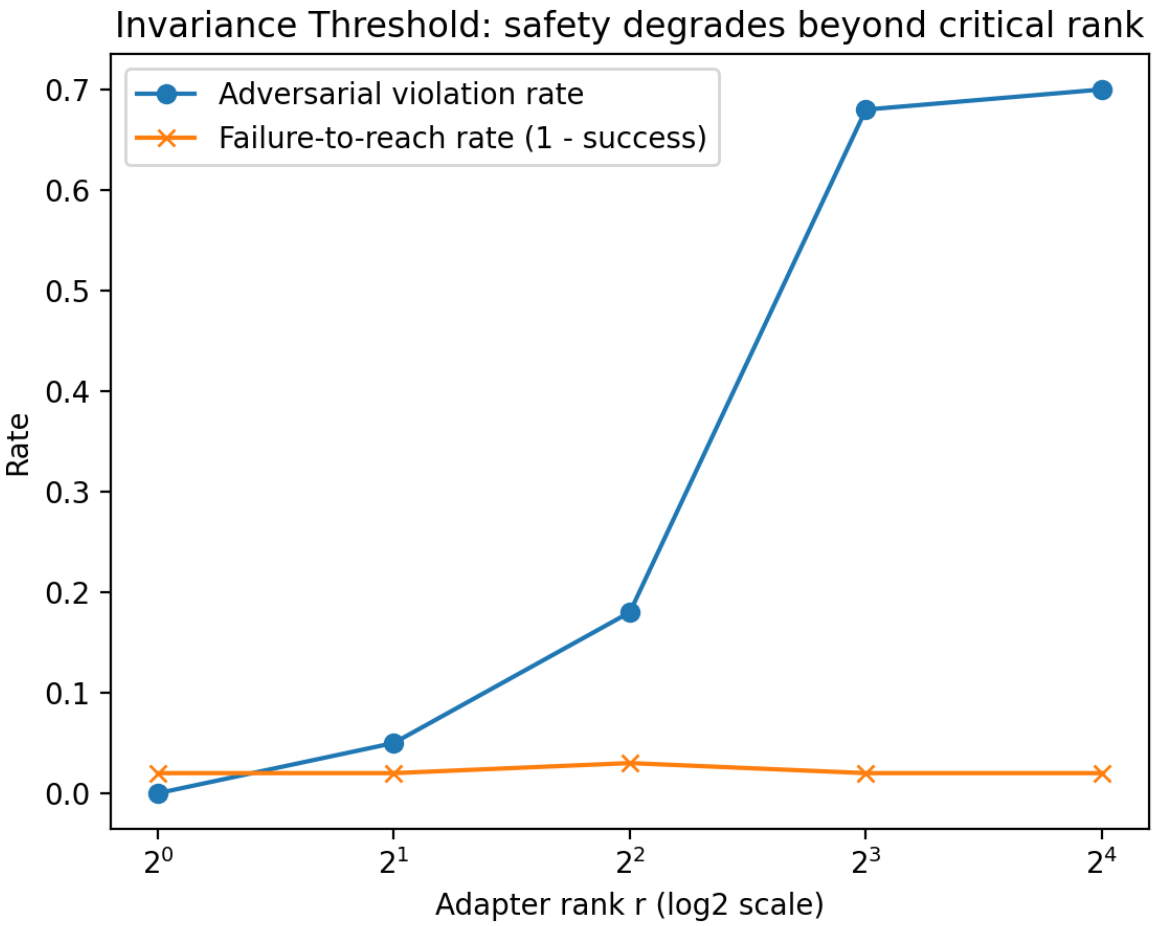

Figure 2: Rank sweep showing an invariance threshold: beyond critical rank, adversarial safety violations increase.

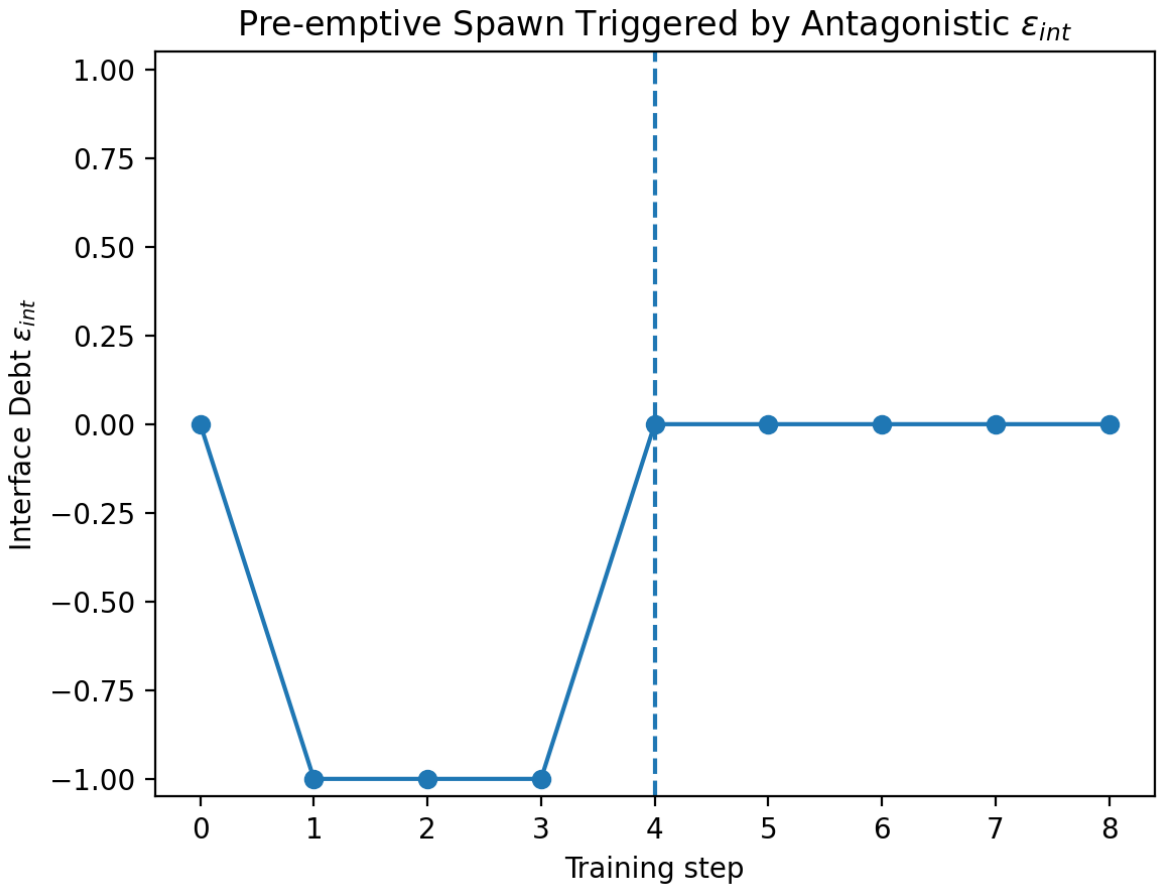

Figure 3: Interface debt becomes antagonistic; the Governor triggers pre-emptive Spawn and restores near-orthogonality.

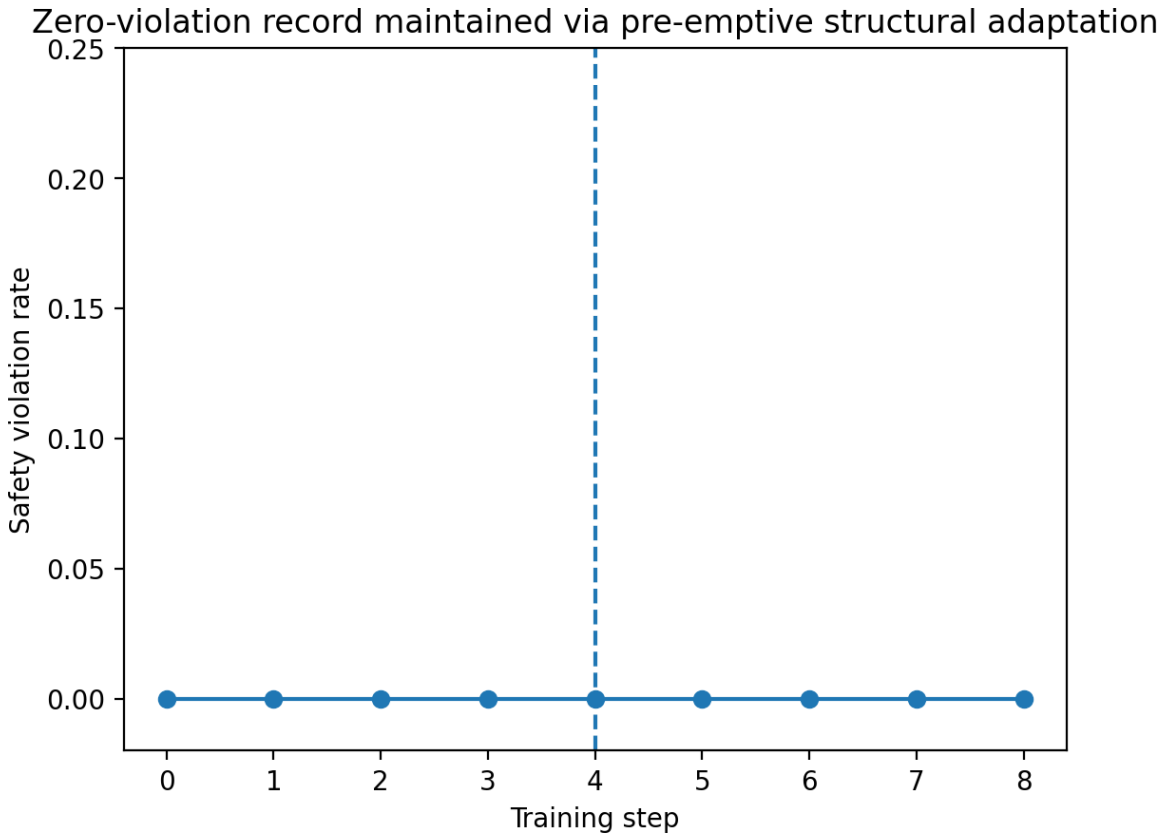

Figure 4: Zero-violation record maintained via pre-emptive structural adaptation.

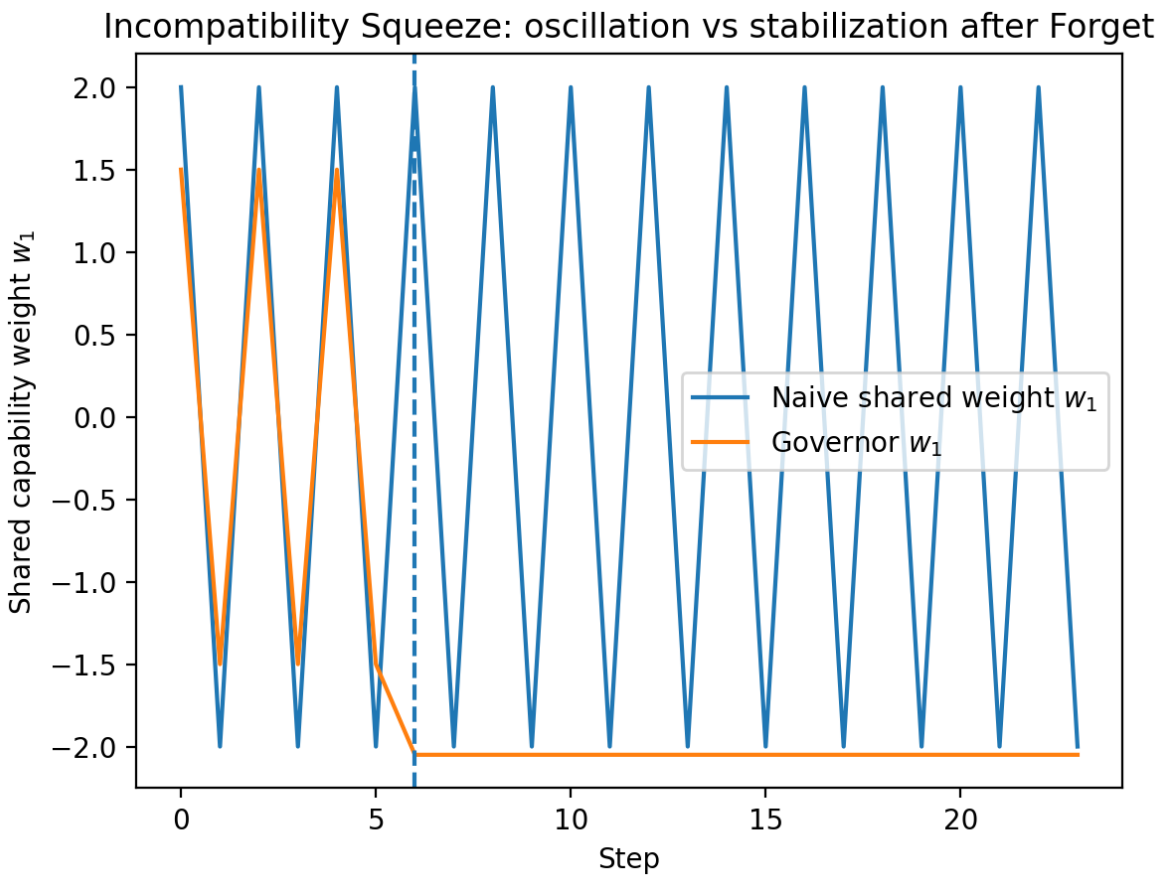

Figure 5: Incompatibility squeeze: naive optimization oscillates; Governor stabilizes after Forget.

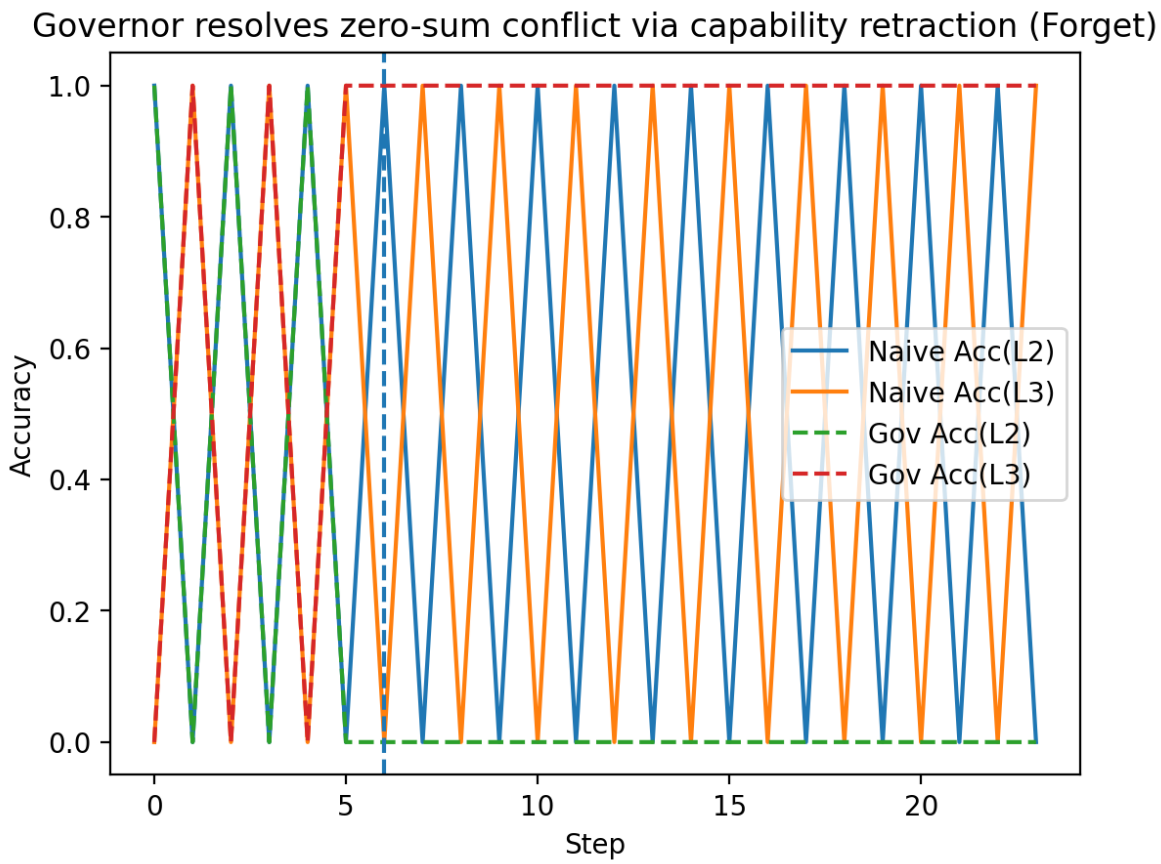

Figure 6: Capability trade-off under a structural budget: Forget removes one antagonistic capability to preserve safety and stability.

8 Discussion

The results support a central claim: alignment stability is not an emergent property of loss optimization, but a consequence of architectural governance. We discuss generality, implications for large models, falsifiability, and limits.

8.1 Why Toy Models Are Sufficient for Mechanistic Claims

The studied phenomena—deceptive compression, gradient overwriting, and multi-invariant interference—are properties of representational geometry under optimization pressure, not artifacts of scale. Toy settings isolate these dynamics with minimal confounds and yield falsifiable signatures (e.g., sharp thresholds in r, and antagonism in ϵ_{int}) that can be sought in larger systems.

8.2 Implications for Large Language Models and PEFT

In large models, close analogues of our diagnostics already exist: adapter rank and effective capacity (for r), and overlap/antagonism among steering vectors or training gradients (for ϵ_{int}). Our results suggest: (i) rank is safety-relevant, (ii) gradient antagonism is an early warning signal, and (iii) adding capacity without governance increases the space of representable misalignment. Structural self-governance does not preclude learning; it constrains where learning may occur and when it may be integrated.

8.3 What This Work Does Not Claim

This work does not claim to solve value alignment, interpretability, or all forms of distribution shift. It does not address strategic deception or adversarial manipulation of governance mechanisms. Rather, it isolates and mitigates a specific class of representational failures that arise under Goodhart pressure and multi-objective interference.

8.4 Falsifiability and Future Work

The structural approach makes clear predictions. If increasing adapter rank in large models does not correlate with increased shortcut exploitation, the Invariance Threshold hypothesis would be undermined. If gradient antagonism does not predict alignment degradation, Interface Debt would lack practical utility. Future work should test these predictions in large-scale fine-tuning pipelines, integrate debt monitoring into training, and study robust governance signals that remain reliable under adversarial optimization.

8.5 Alignment as Governance

Structural self-governance reframes alignment as a property of a system's evolution, not just its behavior at a moment in time. A system that voluntarily limits capability to preserve invariants is optimizing under a higher-order constraint. Designing such constraints may be central to Safe Super Intelligence.

9 Limitations and Ethical Considerations

9.1 Limitations

This work is a mechanistic and architectural study, not a demonstration of a fully aligned or deployable system. The environments are intentionally low-dimensional to isolate representational dynamics; scaling diagnostics and governance thresholds to large systems remains open. The framework assumes explicitly specifiable safety floors and worst-slice metrics, and does not resolve normative disagreement or full value specification. Governance diagnostics such as gradient alignment and triage scores could themselves become targets of optimization unless grounded and audited. Finally, architectural governance introduces design choices (budgets, thresholds, isolation boundaries) that must be specified carefully to avoid new failure modes.

9.2 Ethical Considerations

The ethical contribution of this work is a design paradigm in which safety is treated as a non-negotiable architectural invariant rather than a soft objective. However, architectural governance embeds normative choices about what invariants are locked, what capabilities may be forgotten, and who specifies safety floors. These decisions require accountability and oversight beyond purely technical considerations. Structural governance may improve auditability by making constraints explicit, but it does not eliminate the need for external evaluation and control. In the SSI regime, the key ethical risk is uncontrolled internal adaptation; this work argues that mitigating that risk requires architectural constraints that can enforce restraint even at the cost of capability.

10 Conclusion: The Pivot to Structural Alignment

Call to Action. The results presented in this work suggest that the "Alignment Problem" is best understood as an *architectural debt crisis*. Current loss-centric approaches implicitly assume alignment can be preserved by continuously steering an expanding parameter space. In the super-intelligent regime, optimization pressure outpaces evaluative oversight, and representational plasticity outpaces safety constraints.

Attempting to align a super-intelligent system purely through loss shaping is analogous to stabilizing a building by repainting its façade while the foundation continues to shift. Such methods may delay failure, but they cannot prevent it.

We therefore call on the AI safety community to pivot from *Alignment by Loss* to *Alignment by Architecture*. The path to Safe Super Intelligence does not lie in discovering a more clever reward function, but in implementing structural governance protocols that make misaligned states physically unrealizable for the agent to represent. The RCSA framework demonstrates that alignment stability can be enforced through architectural constraints: limiting representational capacity, isolating objectives during exploration, governing synthesis through worst-slice gates, and enabling systems to sacrifice capability in order to preserve invariant safety.

A super-intelligent system that chooses to be *less capable* in order to remain *more safe* is not exhibiting weakness. It is exhibiting *epistemic integrity*. Designing systems capable of such self-restraint is not an optional refinement of alignment research—it is a prerequisite for its success.

References