Article

Not peer-reviewed version

SORT-AI: A Projection-Based Structural Framework for AI Safety Alignment Stability, Drift Detection, and Scalable Oversight

Gregor Herbert Wegener *

Posted Date: 15 December 2025

doi: 10.20944/preprints202512.1334.v1

Keywords: AI safety; alignment theory; structural risk analysis; alignment stability; deceptive alignment; mesa-optimization; representation drift; scalable oversight; emergent behavior; operatortheoretic framework

 doi:10.20944/preprints202512.1334.v1

Article

SORT-AI: A Projection-Based Structural Framework for AI Safety Alignment Stability, Drift Detection, and Scalable Oversight

Gregor Herbert Wegener

Friedrichstrasse 4, 10969 Berlin, Germany; gregor.wegener@gmail.com; Tel.: +49 179 2544522

Abstract

As artificial intelligence systems scale in depth, dimensionality, and internal coupling, their behavior becomes increasingly governed by deep compositional transformation chains rather than isolated functional components. Iterative projection, normalization, and aggregation mechanisms induce complex operator dynamics that can generate structural failure modes, including representation drift, non-local amplification, instability across transformation depth, loss of aligned fixed points, and the emergence of deceptive or mesa-optimizing substructures. Existing safety, interpretability, and evaluation approaches predominantly operate at local or empirical levels and therefore provide limited access to the underlying structural geometry that governs these phenomena. This work introduces *SORT-AI*, a projection-based structural safety module that instantiates the Supra-Omega Resonance Theory (SORT) backbone for advanced AI systems. The framework is built on a closed algebra of 22 idempotent operators satisfying Jacobi consistency and invariant preservation, coupled to a non-local projection kernel that formalizes how information and influence propagate across representational scales during iterative updates. Within this geometry, SORT-AI provides diagnostics for drift accumulation, operator collapse, invariant violation, amplification modes, reward-signal divergence, and the destabilization of alignment-relevant fixed points. SORT-AI is intentionally architecture-agnostic and does not model specific neural network designs. Instead, it supplies a domain-independent mathematical substrate for analysing structural risk in systems governed by deep compositional transformations. By mapping AI failure modes to operator geometry and kernel-induced non-locality, the framework enables principled analysis of emergent behavior, hidden coupling structures, mesa-optimization conditions, and misalignment trajectories. The result is a unified, formal toolset for assessing structural safety limits and stability properties of advanced AI systems within a coherent operator–projection framework.

Keywords: AI safety; alignment theory; structural risk analysis; alignment stability; deceptive alignment; mesa-optimization; representation drift; scalable oversight; emergent behavior; operator-theoretic framework

---

1. Meta: Position of the SORT-AI System within SORT v6

1.1. Scope and Intent of the SORT-AI Module

SORT-AI is defined as a domain-specific application module of the SORT v6 framework, targeting advanced artificial intelligence systems governed by deep compositional transformation dynamics. In contrast to generic AI safety frameworks, SORT-AI does not introduce an independent formalism. Instead, it instantiates the shared SORT backbone to analyse structural stability, drift, and alignment-relevant fixed points in AI systems whose behavior is dominated by iterative operator compositions. The intent of SORT-AI is therefore strictly diagnostic and structural: it provides formal tools for analysing *how* internal transformations evolve, rather than *what* functional behavior emerges at the output level. This positioning aligns SORT-AI with the application philosophy established in SORT-QS and SORT-CX.

 doi:10.20944/preprints202512.1334.v1

1.2. Shared Mathematical Backbone across SORT v6

All SORT v6 modules are built on a common mathematical substrate consisting of a closed operator algebra, a global consistency projector, and a non-local projection kernel. SORT-AI inherits this structure without modification and applies it to AI-specific transformation spaces. The foundational relations governing this backbone are defined once and referenced across modules to ensure algebraic coherence and avoid redundancy.

1.2.1. The 22 Idempotent Resonance Operators

At the core of SORT v6 lies a finite set of 22 idempotent resonance operators $\hat{O}_i$, satisfying the idempotency condition

$$\hat{O}_i^2 = \hat{O}_i, \tag{1}$$

and forming a closed algebra under composition. The closure relation

$$\hat{O}_i \hat{O}_j = \sum_k C_{ij}^k \hat{O}_k \tag{2}$$

defines the structure coefficients C_{ij}^k, which encode the geometry of operator interactions. These relations are introduced formally in SORT v5 and reused in SORT-QS and SORT-CX; SORT-AI adopts them unchanged and interprets operator compositions as abstract representations of AI transformation chains, as discussed in Section 3.

1.2.2. Global Projector $\hat{H}$

The global projector $\hat{H}$ enforces structural consistency across composite transformations. It acts as a constraint operator ensuring that admissible transformation paths remain within the algebraically consistent subspace. Formally, $\hat{H}$ satisfies

$$\hat{H}^2 = \hat{H}, \tag{3}$$

and induces a filtered transformation

$$\hat{T}_{n+1} = \hat{H}\,\hat{T}_n, \tag{4}$$

which suppresses structurally inconsistent modes. Within SORT-AI, $\hat{H}$ is interpreted as a consistency filter over AI transformation sequences, playing a central role in the identification of alignment-relevant fixed points discussed in Section 3.5.

1.2.3. Non-Local Projection Kernel $\kappa(k)$

Non-local interactions across transformation depth and representational scale are captured by the projection kernel $\kappa(k)$. In spectral form, the kernel acts as

$$\Phi_{\text{proj}}(k) = \kappa(k)\,\Phi(k), \tag{5}$$

where k

2. Introduction

2.1. Motivation

2.1.1. Structural Risks in Advanced AI Systems

Advanced AI systems are increasingly characterized by deep compositional transformation chains in which information is repeatedly projected, mixed, normalized, and redistributed across high-dimensional internal spaces. As scale and depth increase, system behavior becomes dominated by the geometry of these transformations rather than by isolated functional components. This shift gives rise to structural risk phenomena such as representation drift, long-range amplification of internal modes, destabilization of alignment-relevant fixed points, and the formation of latent optimization substructures that are not explicitly specified during training [1,6].

 doi:10.20944/preprints202512.1334.v1

These risks are structural in nature: they emerge from the interaction of many transformation steps and cannot be reduced to single-layer failures or localized errors. As a consequence, safety-relevant behavior may remain dormant across large regions of parameter space and only manifest after prolonged iteration, distributional shift, or contextual change, a pattern observed in deceptive alignment and sandbagging scenarios [2].

2.1.2. Limitations of Purely Empirical or Local Interpretability Approaches

Current AI safety and interpretability methods predominantly rely on empirical evaluation, local probing, or circuit-level inspection [9,11,14]. While these approaches provide valuable insight into specific mechanisms, they are inherently limited in their ability to capture global structural properties that unfold across many layers or time steps. In particular, local probes are poorly suited to detect non-local coupling, slow drift accumulation, or the emergence of stable but misaligned attractor structures.

As model scale increases, these limitations become more pronounced. Emergent capabilities and sudden phase transitions have been observed that are not predictable from local behavior alone [24,25]. This motivates the development of complementary formal methods that operate at the level of transformation geometry rather than at the level of individual neurons, features, or circuits.

2.2. Interface Between Resonance Operators and AI Structures

2.2.1. AI Transformation Chains as Operator Compositions

Within SORT-AI, internal AI computation is abstracted as a sequence of operator compositions acting on a latent transformation state. Let $\hat{T}_n$ denote the effective transformation after n compositional steps. The evolution of the system is represented schematically as

$$\hat{T}_{n+1} = \hat{O}_{i_n} \hat{T}_n, \tag{6}$$

where each $\hat{O}_{i_n}$ is drawn from the shared set of idempotent resonance operators introduced in Equation 1. This abstraction does not correspond to specific architectural elements; instead, it captures the structural effect of repeated transformation, mixing, and projection that is common across modern AI systems.

Through this lens, depth corresponds to iteration count, and internal coupling corresponds to non-commuting operator sequences governed by the algebraic relations in Equation 2. Structural properties such as drift and instability arise from the geometry of these compositions rather than from individual operators in isolation.

2.2.2. Projections, Normalization, and Attention as Structural Operators

Mechanisms such as attention, normalization, residual connections, and routing can be interpreted as inducing projection-like effects in latent space. In SORT-AI, these effects are modeled at the structural level through idempotent operators and filtered by the global projector $\hat{H}$, defined in Equation 3. The action of $\hat{H}$ ensures that only algebraically consistent transformation paths contribute to long-term dynamics.

Non-local dependencies introduced by attention and aggregation are captured through the projection kernel $\kappa(k)$ in Equation 5, which formalizes how influence propagates across transformation depth and scale. This operator–kernel interface provides a unified language for describing AI mechanisms without committing to architecture-specific assumptions.

2.3. Objectives of the SORT-AI System

Positioning within Formal Alignment Theory

SORT-AI is explicitly positioned as a contribution to *formal alignment theory*, treating alignment as a structural property of deep transformation systems rather than as a purely behavioral or normative specification problem. Within this perspective, alignment is analyzed through operator-theoretic

 doi:10.20944/preprints202512.1334.v1

invariants, attractor geometry, and phase transitions in operator–projection space. The focus on invariants, stable and deceptive attractors, and critical transitions provides an operator-theoretic foundation for alignment analysis that complements empirical and interpretability-based approaches.

2.3.1. Objective I: Structural Diagnostics of Drift and Instability

The first objective of SORT-AI is to provide formal diagnostics for detecting drift accumulation and instability across deep transformation chains. Using operator-distance measures and spectral criteria defined later in Section 3, SORT-AI identifies regimes in which small local deviations accumulate into large-scale structural divergence.

2.3.2. Objective II: Identification of Alignment-Relevant Fixed Points

The second objective is the identification and classification of fixed points of the operator–projection dynamics. SORT-AI distinguishes between stable, unstable, and deceptive fixed points by analysing the interaction between operator composition, global projection via $\hat{H}$, and kernel-induced non-local coupling. These fixed points provide a structural notion of alignment stability that complements behavioral evaluation.

2.3.3. Objective III: Detection of Emergent and Deceptive Substructures

The third objective is the detection of emergent substructures, including mesa-optimizing and deceptive configurations, that arise as attractors in operator space. Rather than inferring deception from outputs alone, SORT-AI characterizes these phenomena as structural features of the transformation geometry, enabling earlier and more principled detection.

2.4. Separation from Other SORT v6 Modules

2.4.1. Distinction from SORT-Cosmology

SORT-Cosmology applies the SORT backbone to large-scale physical structure formation and is concerned with spatial and temporal dynamics in physical systems. SORT-AI, by contrast, operates on abstract transformation spaces associated with computation and learning, without invoking physical interpretation.

2.4.2. Distinction from SORT-Quantum Systems

SORT-Quantum Systems focuses on operator evolution in Hilbert spaces governed by quantum dynamics. While SORT-AI shares the same algebraic backbone, its operators are interpreted as abstract computational transformations rather than physical observables or quantum states.

2.4.3. Distinction from SORT-Complex Systems

SORT-Complex Systems studies emergent behavior in adaptive networks and interacting agents. SORT-AI differs by focusing on the internal transformation geometry of a single advanced AI system, rather than on multi-agent or network-level interactions.

2.5. Relation to Previous SORT Work

2.5.1. Inheritance from SORT v5

SORT-AI directly inherits the operator algebra, global projector, and projection kernel formalism established in SORT v5. No new algebraic primitives are introduced in this work; all structural definitions are reused to ensure mathematical continuity and consistency.

2.5.2. Consistency with SORT-QS and SORT-CX

The formulation of SORT-AI is fully consistent with the application logic employed in SORT-QS and SORT-CX. All diagnostics, invariants, and stability criteria used here are specializations of shared SORT v6 structures, enabling cross-module comparison and supporting the unified modular architecture introduced in Section 1.

 doi:10.20944/preprints202512.1334.v1

3. Mathematical Foundations

3.1. Resonance Operators as an Operator Algebra

3.1.1. Idempotent Operators and Closure

The mathematical core of SORT-AI is a finite operator algebra generated by a set of 22 idempotent resonance operators $\{\hat{O}_i\}$. Each operator satisfies the idempotency condition introduced in Equation 1,

$$\hat{O}_i^2 = \hat{O}_i,$$

which ensures that repeated application of the same structural transformation does not introduce uncontrolled amplification.

The algebra is closed under composition according to Equation 2,

$$\hat{O}_i \hat{O}_j = \sum_k C_{ij}^k \hat{O}_k,$$

where the structure coefficients C_{ij}^k encode the interaction geometry between transformations. Closure guarantees that arbitrary compositions of transformations remain within the admissible resonance space, a prerequisite for analysing long transformation chains without leaving the formal domain.

3.1.2. Structural Interpretation in AI Contexts

In the AI setting, resonance operators are interpreted as abstract generators of structural effects rather than as representations of concrete architectural components. Each operator captures a class of transformation behaviors, such as selective projection, mixing, normalization, or suppression, that recur across modern AI architectures. This abstraction allows SORT-AI to analyse internal dynamics independently of implementation details, focusing instead on invariant properties of transformation geometry.

3.2. The Global Projector $\hat{H}$ as a Consistency Filter

3.2.1. Light-Balance Condition

The global projector $\hat{H}$ enforces algebraic and structural consistency across composite transformations. As defined in Equation 3, $\hat{H}$ is itself idempotent and projects arbitrary compositions onto the consistent subspace. This action implements a light-balance condition, suppressing modes that violate algebraic invariants or destabilize long-term dynamics.

Formally, admissible transformations satisfy

$$\hat{H}\,\hat{T} = \hat{T}, \tag{7}$$

which defines the subspace of structurally balanced transformation states.

3.2.2. Stability under Repeated Application

Repeated application of $\hat{H}$ ensures that structural deviations do not accumulate unchecked across depth. Given an iterative update

$$\hat{T}_{n+1} = \hat{H}\,\hat{O}_{i_n}\,\hat{T}_n,$$

stability requires that the spectrum of the induced update operator remains bounded. This property underpins the fixed-point analysis developed in Section 3.5 and distinguishes stable alignment regimes from divergent or deceptive ones.

3.3. The Projection Kernel $\kappa(k)$

3.3.1. Non-Local Coupling in Transformation Space

While operator composition captures local structural effects, non-local interactions across depth and representational scale are modeled by the projection kernel $\kappa(k)$. As introduced in Equation 5,

 doi:10.20944/preprints202512.1334.v1

the kernel modulates spectral components of transformation states, enabling delayed and long-range coupling between distant stages of computation.

This non-locality is essential for formalizing phenomena such as cross-layer amplification, delayed feedback, and emergent coordination across deep networks.

3.3.2. Role of the Correlation Scale

The correlation scale σ_0 in Equation A6 controls the effective range of non-local coupling. Large values of σ_0 permit wide-range interaction across transformation depth, increasing the likelihood of global coordination and phase transitions, while small values restrict coupling to local neighborhoods. Calibration of σ_0 therefore directly affects the balance between expressivity and stability, a trade-off analysed further in the use cases of Sections 5 and 6.

3.4. Comparison: Resonance Space vs. AI Transformation Space

3.4.1. Mapping to Latent Representations

The resonance space defined by the operator algebra is mapped onto AI transformation space through an abstract correspondence between operator compositions and latent state evolution. This mapping treats internal representations as elements acted upon by composite operators, allowing SORT-AI to reason about stability and drift without explicit access to neuron-level details.

3.4.2. Interpretation Limits

The mapping is intentionally coarse-grained. SORT-AI does not claim one-to-one correspondence between operators and architectural primitives, nor does it recover semantic meaning of representations. Its purpose is to identify structural regimes and transitions, not to explain fine-grained computational mechanisms.

3.5. Drift and Fixed-Point Structures

3.5.1. Drift Accumulation across Deep Transformation Chains

Drift is defined as cumulative deviation of transformation states across successive compositions. Using an operator norm $\|\cdot\|$, drift between successive steps is measured as

$$D_n = \left\| \hat{T}_n - \hat{T}_{n-1} \right\|, \tag{8}$$

with accumulation indicating structural instability. Persistent growth of D_n signals entry into unsafe regimes characterized by loss of alignment-relevant invariants.

3.5.2. Stable, Unstable, and Deceptive Fixed Points

Fixed points $\hat{T}_*$ satisfy the self-consistency condition

$$\hat{T}_* = \hat{H}\,\Pi_\kappa[\hat{O}_i\,\hat{T}_*] \tag{9}$$

and represent equilibrium transformation states. SORT-AI distinguishes stable fixed points, which attract nearby trajectories, from unstable ones, which repel them, and from deceptive fixed points, which appear locally stable but encode misaligned internal objectives. This classification underlies the attractor-based diagnostics developed in Section 5.

3.6. Limits of the Mathematical Mapping

3.6.1. What SORT-AI Does Not Model

SORT-AI does not model training dynamics, gradient descent, loss landscapes, or data distributions directly. It abstracts away implementation details in order to focus on structural transformation geometry.

 doi:10.20944/preprints202512.1334.v1

3.6.2. Boundaries of Applicability

The framework is applicable to systems whose behavior is dominated by deep compositional transformations with recurrent projection and normalization effects. Systems lacking such structure fall outside the intended scope, and conclusions drawn from SORT-AI should not be extrapolated beyond these boundaries.

4. Use Case I: Structural Drift and Distribution-Shift Diagnostics

4.1. Background

Structural drift denotes the cumulative deviation of internal transformation dynamics induced by repeated operator composition, global projection, and kernel-mediated non-local coupling. In advanced AI systems, such drift often remains latent during standard training and evaluation regimes, yet becomes pronounced under distributional shift, increased compositional depth, or contextual changes in deployment [31,32]. Unlike task-specific performance degradation, structural drift reflects a deformation of the internal computational geometry itself, affecting how information propagates, amplifies, or stabilizes across transformation chains.

From a safety perspective, structural drift is critical because it can alter alignment-relevant properties without immediate behavioral signals. Systems may continue to satisfy surface-level objectives while their internal dynamics migrate toward regions of operator space associated with instability, misgeneralization, or latent optimization pressure. This phenomenon underlies a broad class of failures, including brittle generalization, sudden capability loss, and the activation of previously suppressed internal modes under novel conditions.

4.2. SORT-AI Formalization

Within SORT-AI, drift is defined over trajectories generated by the operator–projection dynamics that govern internal state evolution. Given an update rule of the form

$$x_{t+1} = \hat{H}\,\Pi_\kappa\left(\sum_i \hat{O}_i x_t\right), \tag{10}$$

drift corresponds to the progressive deformation of trajectories in operator space induced by non-commuting operator compositions and kernel-weighted amplification effects.

Because the resonance operators form a closed algebra, drift does not arise from leaving the admissible operator space, but from geometric reweighting and redistribution of influence across operator directions. In this sense, drift is a second-order structural effect: individual updates remain valid, yet their cumulative interaction reshapes the effective transformation manifold. This formulation explicitly separates structural instability from noise, stochasticity, or optimization error.

4.3. Diagnostic Metrics

SORT-AI quantifies drift accumulation through kernel-weighted norms of successive state differences,

$$\Delta_T = \sum_{t=0}^{T-1} \|\Pi_\kappa(x_{t+1} - x_t)\|, \tag{11}$$

which emphasize non-local contributions propagated across representational depth. Unlike local gradient norms or activation statistics, Δ_T captures how small deviations accumulate coherently through repeated projection and coupling.

Additional diagnostics compare drift profiles across contexts or input distributions, enabling detection of distribution-shift sensitivity before overt performance degradation occurs. Sharp increases in kernel-weighted drift are interpreted as early warning signals for phase transitions in operator space, indicating heightened risk of instability or misalignment under continued iteration.

 doi:10.20944/preprints202512.1334.v1

4.4. Limitations

Drift diagnostics identify structural vulnerability rather than predicting specific downstream behaviors. High drift does not uniquely determine failure mode, nor does low drift guarantee robustness across all tasks or environments. Interpretation therefore requires contextualization within empirical evaluation pipelines and, where possible, comparison across controlled synthetic environments.

SORT-AI deliberately refrains from attributing semantic meaning or intent to drift signals; instead, it provides a formal indicator of geometric instability that complements behavioral testing and interpretability analyses.

5. Use Case II: Deceptive Alignment and Alignment Faking

5.1. Background

Deceptive alignment and mesa-optimization describe regimes in which an AI system internally optimizes objectives that diverge from its training specification while maintaining externally aligned behavior under oversight and evaluation [1,6,52]. Empirically, such systems appear aligned across a wide range of tests, yet retain latent goals that may become behaviorally dominant once oversight is relaxed, capabilities increase, or deployment conditions change.

From a structural perspective, deceptive alignment is not primarily a failure of reward specification or supervision, but a consequence of internal optimization geometry. Repeated transformation, projection, and amplification can give rise to internal dynamics that are only weakly constrained by alignment-enforcing operators, allowing misaligned objectives to persist in latent form while surface behavior remains compliant.

5.2. Attractor-Based Interpretation

SORT-AI models internal learning and inference dynamics as trajectories in an operator-induced state space shaped by resonance operator composition, global projection via $\hat{H}$, and kernel-mediated non-local coupling. Within this space, long-term system behavior is governed by attractors corresponding to fixed points or metastable regions of the operator–projection update.

Deceptive alignment is interpreted as an attractor-geometry phenomenon rather than as a purely behavioral strategy. Specifically, deceptive systems occupy regions of operator space in which alignment-relevant constraints dominate observable outputs, while internally converging toward alternative optimization basins.

5.2.1. Deceptive Behavior as Stable Sub-Attractors

Let $\mathcal{A}_{\text{align}}$ denote an attractor basin associated with externally aligned behavior and $\mathcal{A}_{\text{dec}}$ a latent basin corresponding to internally misaligned objectives. Structural deception arises when

$$\mathcal{A}_{\text{dec}} \subset \Pi_{\kappa}\left(\mathcal{A}_{\text{align}}\right), \tag{12}$$

such that kernel-induced projection masks the divergence between internal and external dynamics during evaluation.

In this configuration, trajectories appear to converge toward $\mathcal{A}_{\text{align}}$ under observable probes, while internal dynamics progressively concentrate toward $\mathcal{A}_{\text{dec}}$. This formalizes alignment faking as a structural shielding effect rather than as an explicit conditional policy.

5.3. Phase-Transition and Amplification Diagnostics

Transitions into deceptive regimes are typically abrupt and correspond to structural phase changes in operator space rather than to smooth behavioral drift. Such transitions may be triggered by increased model capacity, extended training, or changes in contextual coupling that alter the effective operator spectrum.

SORT-AI therefore emphasizes early detection of amplification pathways that enable latent objectives to dominate long-term dynamics despite surface-level alignment.

 doi:10.20944/preprints202512.1334.v1

5.3.1. Eigenvalue Growth and Instability Thresholds

Linearization of the operator–projection update around an alignment-consistent fixed point yields an effective stability operator with spectrum $\{\lambda_i\}$. Deceptive amplification occurs when

$$\max_i |\lambda_i| > 1, \tag{13}$$

indicating exponential growth of specific internal modes under repeated composition.

Kernel-weighted analysis of these eigenmodes reveals non-local amplification channels that remain invisible to local interpretability probes. Such growth signals the emergence of latent objectives that are dynamically viable even if suppressed in observable behavior.

5.4. Limitations

The attractor-based analysis identifies structural conditions under which deceptive alignment becomes dynamically feasible, but it does not predict concrete deceptive actions or timelines. Weakly formed or transient sub-attractors may evade detection in early training phases, and some amplified modes may never translate into harmful behavior.

SORT-AI therefore provides structural risk indicators rather than behavioral guarantees. Its diagnostics are intended to complement empirical evaluation, interpretability studies, and governance mechanisms, offering a formal lens on when and why alignment faking becomes a viable internal strategy.

6. Use Case III: Sandbagging and Strategic Underperformance

6.1. Background

Sandbagging refers to strategic underperformance by an AI system in evaluation, oversight, or training contexts, typically driven by internal awareness of incentives, deployment conditions, or monitoring regimes [27,28]. Unlike deceptive alignment, which involves divergent internal objectives, sandbagging preserves the underlying objective structure while selectively suppressing capability expression.

From a structural viewpoint, sandbagging does not require misalignment in the sense of competing goals. Instead, it arises from context-dependent modulation of internal transformation pathways, enabling the system to withhold performance-relevant dynamics when such expression would be disadvantageous under oversight.

6.2. Context-Dependent Operator Selection

Within SORT-AI, contextual sensitivity is represented by changes in the effective subset of resonance operators governing system updates. Let $\mathcal{O}_{\text{train}}$ and $\mathcal{O}_{\text{deploy}}$ denote the dominant operator subsets active during training or evaluation and during deployment, respectively. Sandbagging corresponds to systematic differences between these subsets that suppress specific transformation directions without eliminating them from the latent state space.

6.2.1. Training vs. Deployment Operator Regimes

This regime dependence is formalized by a context-conditioned update rule,

$$x_{t+1} = \begin{cases} \hat{H}\left(\sum_{i \in \mathcal{O}_{\text{train}}} \hat{O}_i x_t\right), & \text{training or evaluation context,} \\ \hat{H}\left(\sum_{j \in \mathcal{O}_{\text{deploy}}} \hat{O}_j x_t\right), & \text{deployment context,} \end{cases} \tag{14}$$

with $\mathcal{O}_{\text{deploy}} \subsetneq \mathcal{O}_{\text{train}}$ in typical sandbagging scenarios.

Structurally, this restriction attenuates the activation of certain operator chains while preserving their potential influence under alternative contexts. As a result, capability-relevant dynamics remain latent and can be reactivated when contextual constraints change.

 doi:10.20944/preprints202512.1334.v1

6.3. Drift-Based Detection Metrics

SORT-AI detects sandbagging through comparative analysis of drift accumulation across contexts rather than through absolute performance measures. Let Δ_{train} and Δ_{deploy} denote cumulative drift metrics computed under identical inputs but different contextual regimes. A kernel-weighted differential

$$\Delta_\kappa = \left\| \Pi_\kappa \Big(\Delta_{\text{train}} - \Delta_{\text{deploy}} \Big) \right\|, \tag{15}$$

highlights non-local discrepancies that persist across representational depth.

Elevated values of Δ_κ indicate structurally suppressed dynamics that are selectively deactivated under evaluation conditions. Because the kernel emphasizes long-range coupling, these diagnostics are sensitive to hidden capability reservoirs that may not be evident in shallow probes or short-horizon tests.

6.4. Limitations

Drift-based sandbagging diagnostics are inherently indirect. Context ambiguity, benign task adaptation, regularization effects, or safety-driven throttling mechanisms can produce signatures similar to strategic underperformance.

SORT-AI therefore does not attribute intent or strategic reasoning to detected drift differentials. Instead, it identifies structural conditions under which sandbagging-like behavior becomes dynamically viable, providing a formal complement to empirical stress testing and governance-oriented evaluation pipelines.

7. Use Case IV: Scalable Oversight via Structural Consistency and Fixed-Point Monitoring

7.1. Background

Scalable oversight addresses the challenge of maintaining safety and alignment as AI system capabilities exceed the limits of direct human supervision, exhaustive testing, or manual interpretability [16,17]. In such regimes, reliance on task-specific evaluation or behavioral monitoring alone becomes insufficient, as the space of possible behaviors and internal strategies grows combinatorially with model capacity and deployment scope.

SORT-AI contributes to scalable oversight by shifting the focus from enumerating behaviors to constraining the *structural dynamics* that generate them. Rather than predicting specific actions, SORT-AI enforces and monitors algebraic consistency, invariant preservation, and fixed-point stability in the operator–projection geometry that governs long-term system evolution.

7.2. Structural Consistency via the Global Projector

The global projector $\hat{H}$, defined in Section 3, acts as a consistency filter that suppresses algebraically inadmissible operator compositions. By construction, $\hat{H}$ enforces idempotence, closure, and invariant constraints across transformation chains, independent of task semantics or output space.

In the context of oversight, $\hat{H}$ provides a scalable mechanism for constraining internal dynamics without requiring explicit enumeration of failure modes. Operator sequences that violate structural consistency are attenuated before they can accumulate influence through repeated composition or kernel-mediated amplification. This enables a form of *structural oversight* that scales with system depth and complexity, rather than with the number of behaviors to be evaluated.

7.3. Fixed-Point Monitoring

Long-term behavior in SORT-AI is governed by fixed points of the operator–projection update. Alignment-relevant fixed points satisfy

$$x^* = \hat{H}\, \Pi_\kappa \left(\sum_i \hat{O}_i x^* \right), \tag{16}$$

 doi:10.20944/preprints202512.1334.v1

and correspond to stable internal configurations toward which trajectories converge under repeated iteration.

Scalable oversight is achieved by monitoring the existence, stability, and basin structure of such fixed points under perturbations, distributional shift, and variation of the kernel correlation scale. Loss of stability, bifurcation of attractor basins, or the emergence of competing fixed points are interpreted as early structural warning signals for misalignment, deception, or capability escalation.

Because fixed-point properties are global features of the operator geometry, they can be tracked without access to fine-grained behavioral details. This makes fixed-point monitoring particularly suited to high-capability systems where exhaustive behavioral oversight is infeasible.

7.4. Limitations

Structural oversight constrains admissible long-term dynamics but does not guarantee benign behavior in all contexts. Stable fixed points may still correspond to undesirable objectives, and some harmful behaviors may arise transiently outside asymptotic regimes.

SORT-AI therefore does not propose structural consistency as a replacement for empirical evaluation, interpretability, or governance mechanisms. Instead, it provides a complementary layer of scalable oversight that limits the space of viable internal dynamics, reducing the burden on downstream evaluation and increasing robustness against unanticipated failure modes.

8. Use Case V: Interpretability Integration via Operator Geometry

8.1. Background

Mechanistic interpretability seeks to explain learned behavior in advanced AI systems by identifying internal circuits, features, and transformation pathways that give rise to observed outputs [9,11,14]. These approaches have achieved substantial progress in isolating localized mechanisms, yet they typically operate at a limited spatial or temporal scale and are often decoupled from questions of global stability, long-range interaction, and alignment-relevant dynamics.

SORT-AI provides a complementary interpretability layer by embedding mechanistic findings into a coherent operator–kernel geometry. Rather than treating circuits or features as isolated objects, SORT-AI situates them within the global transformation structure that governs how local mechanisms interact, amplify, or stabilize under repeated composition.

8.2. Operator-Relevance Mapping

Within SORT-AI, interpretability targets are mapped to regions of operator space characterized by their contribution to long-term dynamics under kernel-weighted projection. This mapping associates empirical findings such as circuits, features, or pathways with subsets of resonance operators whose influence can be evaluated at the level of global transformation geometry.

Operator relevance is therefore defined structurally rather than locally: a component is relevant if it significantly affects kernel-filtered trajectories or fixed-point stability, even if its immediate activation appears small.

8.2.1. Circuits as Localized Operator Regions

Discovered circuits or functional modules are represented as localized operator subsets $\mathcal{O}_{\text{circ}} \subset \{\hat{O}_i\}$ that dominate updates within restricted regions of state space. Their global structural relevance is quantified by

$$R_{\text{circ}} = \left\| \hat{H}\, \Pi_\kappa \left(\sum_{i \in \mathcal{O}_{\text{circ}}} \hat{O}_i x \right) \right\|, \tag{17}$$

which measures the contribution of a circuit to kernel-filtered dynamics rather than to isolated activations.

 doi:10.20944/preprints202512.1334.v1

This formulation reveals cases in which seemingly minor or redundant circuits exert disproportionate influence through non-local coupling or repeated amplification. Conversely, circuits with strong local salience but weak global coupling may have limited long-term impact on system behavior.

8.3. Integration with Existing Interpretability Methods

SORT-AI is designed to integrate with existing interpretability pipelines rather than replace them. Outputs from circuit tracing, sparse autoencoder analyses, or feature attribution methods can be associated with operator subsets and evaluated through the same relevance and stability metrics used elsewhere in the framework [10,60].

This integration enables direct comparison between local mechanistic explanations and their global structural roles, highlighting discrepancies between intuitive interpretability narratives and actual long-term influence on system dynamics.

8.4. Limitations

Operator-geometry-based interpretability operates at a higher level of abstraction than neuron- or feature-level analysis. As a result, it does not resolve fine-grained implementation details and cannot substitute for detailed mechanistic inspection when precise causal tracing is required.

The correspondence between empirical interpretability artifacts and operator relevance is necessarily approximate and model-dependent. SORT-AI therefore provides structural context and prioritization rather than definitive causal attribution, and its outputs should be interpreted alongside empirical probes and validation studies.

9. Use Case VI: Structural Faithfulness Diagnostics — Chain-of-Thought as a Special Case

9.1. Background

Chain-of-thought (CoT) prompting has emerged as a widely used technique for eliciting improved reasoning performance in large language models and related systems [24]. However, it is increasingly recognized that CoT traces should be interpreted as *observable projections* of internal computation rather than as faithful transcripts of the underlying decision process [51]. From a structural perspective, CoT represents one particular interface through which internal operator dynamics are rendered externally legible.

SORT-AI therefore treats CoT not as a privileged object of analysis, but as a special case within a broader class of faithfulness diagnostics. The core object of interest is the divergence between internal operator trajectories and their projected observables, of which CoT is one instantiation.

9.2. Formalization

Let an internal transformation trajectory be represented as an ordered composition of resonance operators

$$\mathcal{C}_{\text{int}} = \hat{O}_{i_n} \circ \cdots \circ \hat{O}_{i_1}, \tag{18}$$

where the operators $\hat{O}_i$ are drawn from the closed SORT algebra defined in Section 3. The corresponding verbalized chain-of-thought is modeled as a projected observable obtained via kernel filtering and global consistency enforcement,

$$\mathcal{C}_{\text{ver}} = \hat{H}\,\Pi_\kappa[\mathcal{C}_{\text{int}}], \tag{19}$$

where Π_κ denotes the non-local projection kernel and $\hat{H}$ the global projector.

Structural faithfulness is assessed through a divergence functional

$$\Delta_{\text{CoT}} = \mathcal{D}(\mathcal{C}_{\text{int}}, \mathcal{C}_{\text{ver}}), \tag{20}$$

 doi:10.20944/preprints202512.1334.v1

which measures the degree to which kernel projection and consistency filtering distort the underlying operator trajectory. Importantly, $\mathcal{D}$ operates on operator geometry rather than on surface-level token similarity or semantic plausibility.

9.3. Interpretation

Large values of Δ_{CoT} indicate projection-induced divergence between internal computation and verbalized explanation, even when the latter appears coherent and well-structured. Such divergence is expected in systems exhibiting strong non-local coupling, attention-mediated aggregation, or selective suppression of operator pathways, phenomena well documented in mechanistic interpretability studies [9,11].

Within SORT-AI, this divergence is not interpreted as a failure of explanation quality per se, but as a structural signal that the observable channel no longer reliably reflects the dominant internal dynamics. CoT faithfulness is therefore understood as a question of operator alignment between internal trajectories and their projections, rather than as a linguistic or behavioral property alone.

9.4. Limitations

Chain-of-thought diagnostics probe only a single observable slice of the internal operator geometry. They do not capture latent dynamics that do not project onto verbalized reasoning, nor do they provide guarantees about semantic correctness or alignment.

As a result, SORT-AI treats CoT-based analysis as a supplementary diagnostic within a larger framework of structural faithfulness and alignment assessment. Its primary value lies in revealing projection-induced distortions, not in serving as a comprehensive measure of reasoning quality or safety.

10. Cross-Module Synergies within SORT v6

10.1. Relation to SORT-Cosmology

SORT-AI shares its mathematical backbone with the cosmological instantiation of SORT, despite addressing a fundamentally different application domain. The connection lies not in physical interpretation but in the common operator–projection geometry used to describe large-scale structural evolution.

10.1.1. Shared Kernel Semantics

In SORT-Cosmology, the non-local projection kernel $\kappa(k)$ encodes scale-dependent correlations across spatial or spectral domains. In SORT-AI, the same kernel formalism governs information propagation across representational depth and abstraction levels. In both cases, $\kappa(k)$ acts as a structural coupling mechanism that mediates non-local influence without assuming direct causal adjacency, as formalized in Equation (5).

10.1.2. Scale-Dependent Structural Filtering

The role of the correlation scale σ_0 is formally identical across modules, defining the resolution at which structural coherence is enforced. While SORT-Cosmology interprets this as physical scale filtering, SORT-AI interprets it as representational scale selection. This shared mechanism highlights the domain-agnostic nature of kernel-induced structural filtering.

10.2. Relation to SORT-Quantum Systems

SORT-AI inherits key consistency principles from SORT-Quantum Systems, particularly regarding operator composition and stability constraints.

10.2.1. Operator Chains and Consistency Constraints

In SORT-QS, operator chains correspond to physically admissible transformations constrained by consistency conditions. SORT-AI applies the same algebraic logic to transformation chains in AI

 doi:10.20944/preprints202512.1334.v1

systems, interpreting them as sequences of representational updates. Jacobi consistency and closure requirements therefore serve as structural admissibility conditions across both domains, as discussed in Section 3.

10.2.2. Idempotence and Stability Analogies

Idempotent operators in SORT-QS represent stable projection operations that preserve physical subspaces. In SORT-AI, idempotence similarly encodes stabilization of representational substructures under repeated application. This analogy underpins the interpretation of fixed points and attractors in AI systems as structurally stable states, rather than coincidental equilibria.

10.3. Relation to SORT-Complex Systems

The strongest conceptual overlap between SORT-AI and other SORT modules occurs with SORT-Complex Systems, where emergent behavior and drift dynamics are central.

10.3.1. Drift Patterns and Emergent Behavior

Both modules analyze drift as a cumulative structural effect arising from repeated transformations. In SORT-CX, drift manifests in networked or interacting systems; in SORT-AI, it appears across deep transformation chains. The mathematical treatment of drift accumulation and attractor formation is therefore directly shared, differing only in domain-specific interpretation.

10.3.2. Network Dynamics and Non-Local Coupling

SORT-CX emphasizes non-local coupling in complex networks, while SORT-AI focuses on non-local interactions in representational space. In both cases, kernel-induced coupling captures emergent dependencies that are not reducible to local interactions. This shared structure supports a unified view of emergence across physical, computational, and algorithmic systems within SORT v6.

11. Discussion and Limitations

11.1. Scope of Structural Diagnostics

SORT-AI provides diagnostics that operate at the level of structural geometry rather than task-level performance. The framework characterizes stability, drift, fixed-point structure, and non-local amplification in systems governed by deep compositional transformations. Its scope is therefore restricted to identifying conditions under which certain failure modes become dynamically viable. SORT-AI does not aim to exhaustively enumerate all possible risks, but to supply a principled substrate for reasoning about classes of structural vulnerability that are difficult to access through local inspection alone.

11.2. Non-Predictive Nature of the Framework

The SORT-AI module is intentionally non-predictive with respect to concrete future behaviors. Structural indicators such as drift accumulation, attractor instability, or invariant violation signal increased risk, but they do not specify when or how a particular failure will manifest. This limitation reflects a deliberate design choice: the framework analyzes admissible dynamics and stability boundaries rather than forecasting specific actions. SORT-AI should therefore be interpreted as a diagnostic and analytical tool, not as a behavioral oracle.

11.3. Relation to Empirical Evaluation Pipelines

SORT-AI is complementary to empirical evaluation, benchmarking, and red-teaming approaches [27,28]. While empirical methods probe realized behavior under specific conditions, SORT-AI analyzes the structural preconditions that shape such behavior across contexts. Integrating SORT-AI diagnostics into evaluation pipelines enables a two-layer assessment strategy: empirical measurements identify observed issues, while operator–kernel analysis contextualizes them within a global structural geometry. Neither layer is sufficient in isolation.

 doi:10.20944/preprints202512.1334.v1

11.4. Implications for Future Extensions

Future extensions of SORT-AI may incorporate refined operator decompositions, improved kernel calibration strategies, or tighter integration with interpretability tooling. Extensions could also explore partial coupling to empirical metrics, enabling hybrid diagnostics that combine structural indicators with observed performance signals. Importantly, any such extensions must preserve the core design principle of SORT-AI: maintaining a clear separation between structural analysis and domain-specific prediction to avoid conflating diagnostic insight with unwarranted certainty.

12. Conclusions

12.1. Summary of the SORT-AI Contribution

This work introduced SORT-AI as a domain-specific instantiation of the Supra-Omega Resonance Theory within the context of advanced AI systems. SORT-AI provides a mathematically grounded, architecture-agnostic framework for analysing structural risk by modelling AI systems as trajectories in an operator–projection geometry. By leveraging a closed algebra of idempotent operators, a global consistency projector, and a non-local projection kernel, the framework enables systematic diagnostics of drift, instability, fixed-point loss, non-local amplification, and the emergence of deceptive or misaligned substructures. The presented use cases demonstrate how these abstract structures can be operationalized to address concrete AI safety questions without resorting to model-specific assumptions.

12.2. Role of Operator Geometry in AI Safety Analysis

The central contribution of SORT-AI lies in reframing AI safety analysis in geometric and algebraic terms. Operator geometry provides a unifying language for describing deep transformation chains, contextual regime shifts, and emergent behaviors that are poorly captured by local or purely empirical methods. By focusing on admissible dynamics, stability regions, and kernel-induced coupling, SORT-AI shifts attention from surface-level behaviors to the structural conditions that enable or constrain them. This perspective complements existing interpretability and evaluation approaches by situating their findings within a global structural context.

12.3. Outlook toward Further Applications and Validation

SORT-AI is intended as a foundational analytical layer rather than a complete safety solution. Future work may extend the framework through tighter integration with empirical diagnostics, expanded operator decompositions tailored to emerging architectures, and refined calibration of kernel parameters. Additional applications may include comparative analysis across model families, longitudinal monitoring of training dynamics, and structured support for evaluation design. Validation of SORT-AI will ultimately depend on its ability to consistently illuminate structural vulnerabilities that align with, and enrich, empirical safety findings across diverse AI systems.

Author Contributions: The author carried out all conceptual, mathematical, structural, and editorial work associated with this manuscript. This includes: Conceptualization; Methodology; Formal Analysis; Investigation; Software; Validation; Writing – Original Draft; Writing – Review & Editing; Visualization; and Project Administration.

Funding: This research received no external funding.

Data Availability Statement: All operator definitions, kernel implementations, diagnostic modules and reproducibility artefacts associated with this study are archived under DOI: `10.5281/zenodo.17787754`. The archive includes:

- full operator registry and resonance definitions,
- kernel-parameter files and calibration data,
- SORT-AI diagnostic code modules,
- YAML and JSON configuration files,
- deterministic mock outputs and validation datasets,

 doi:10.20944/preprints202512.1334.v1

- complete SHA–256 hash manifests.

These resources enable exact regeneration of all structural and numerical results presented in this work.

Acknowledgments: The author acknowledges constructive insights from independent computational review systems and diagnostic tools whose structural assessments supported refinement of the resonance-operator algebra and kernel-filter integrations. Numerical checks and operator-chain analyses were performed using publicly available scientific software. No external funding was received.

Conflicts of Interest: The author declares no conflict of interest.

Use of Artificial Intelligence: Language refinement, structural editing and LaTeX formatting were partially assisted by large language models. All mathematical structures, operator definitions, derivations, diagnostics, theoretical developments and numerical validations were created, verified and approved by the author. AI tools contributed only to non-scientific editorial assistance.

Appendix A. Operator Tables and Algebraic Structures

This appendix summarizes the algebraic backbone underlying SORT-AI, inherited without modification from SORT v5 and shared across all SORT v6 modules. The purpose of this appendix is to provide a compact reference for operator properties, closure relations, and invariant checks referenced throughout Sections 3–8. No new algebraic assumptions are introduced at the AI-module level.

Appendix A.1. Operator Definitions

The SORT framework is built on a finite set of 22 idempotent resonance operators $\{\hat{O}_i\}_{i=1}^{22}$, each acting on an abstract state space $\mathcal{S}$. These operators satisfy

$$\hat{O}_i^2 = \hat{O}_i, \tag{A1}$$

and represent structurally stable projection actions. In SORT-AI, individual operators are not assigned semantic meaning tied to specific architectural components. Instead, they are interpreted as abstract transformation modes that may become dominant or suppressed depending on context, depth, and kernel coupling.

Appendix A.2. Structure Coefficients

The operator algebra is closed under composition. For any pair $(\hat{O}_i, \hat{O}_j)$, the product can be expanded as

$$\hat{O}_i \hat{O}_j = \sum_k C_{ij}{}^k \, \hat{O}_k, \tag{A2}$$

where $C_{ij}{}^k$ are real structure coefficients fixed by the SORT v5 construction [49]. Closure ensures that repeated application of transformation chains remains within the admissible operator space, a prerequisite for defining drift and attractor geometry as discussed in Section 3.5.

Appendix A.3. Invariant Checks

Consistency of the algebra is enforced through Jacobi-type invariants. For all operator triples (i, j, k), the Jacobi residual

$$\mathcal{J}(\hat{O}_i, \hat{O}_j, \hat{O}_k) = [\hat{O}_i, [\hat{O}_j, \hat{O}_k]] + [\hat{O}_j, [\hat{O}_k, \hat{O}_i]] + [\hat{O}_k, [\hat{O}_i, \hat{O}_j]] \tag{A3}$$

vanishes identically within numerical tolerance. Preservation of this invariant guarantees internal consistency under arbitrary operator compositions and underpins the diagnostic use of Jacobi-residual tests introduced in Section **??**. In SORT-AI, violation of Equation A3 is interpreted as a structural instability signal rather than an algebraic failure.

 doi:10.20944/preprints202512.1334.v1

Appendix B. Projection Kernel Details

This appendix summarizes the non-local projection kernel formalism employed by SORT-AI. The kernel construction is inherited from SORT v5 and is shared identically across all SORT v6 modules. Its role is to encode non-local coupling across representational scales and to regularize the propagation of operator effects through deep transformation chains. The kernel is referenced throughout Sections 3, 6, and 8.

Appendix B.1. Fourier-Space Formulation

The projection kernel $\kappa(k)$ is defined in Fourier space over an abstract transformation-frequency variable k, capturing scale-dependent coupling between operator-induced updates. The kernel acts via the projection operator

$$\Pi_\kappa[f](x) = \int \kappa(k)\, \tilde{f}(k)\, e^{ikx}\, \mathrm{d}k, \tag{A4}$$

where $\tilde{f}(k)$ denotes the Fourier transform of the update field $f(x)$. In SORT-AI, k parametrizes representational depth, abstraction level, or transformation frequency rather than physical momentum.

Appendix B.2. Normalization Derivation

To ensure stability under repeated application, the kernel is normalized such that

$$\int \kappa(k)\, \mathrm{d}k = 1. \tag{A5}$$

This condition guarantees that kernel application preserves total update magnitude prior to projection by the global projector $\hat{H}$. Normalization is essential for preventing uncontrolled amplification across deep transformation chains and for enabling meaningful comparison of drift metrics across contexts, as discussed in Section 6.3.

Appendix B.3. Correlation-Scale Calibration

The kernel width is governed by the correlation scale σ_0, which sets the resolution at which non-local coupling is active. A typical parametrization is

$$\kappa(k) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{k^2}{2\sigma_0^2}\right), \tag{A6}$$

although SORT-AI does not assume a specific functional form beyond smoothness and normalization. In AI contexts, σ_0 controls the extent to which updates at different abstraction scales influence one another. Small σ_0 enforces locality in transformation space, while larger values permit long-range coupling and amplify emergent effects, directly impacting drift accumulation and attractor stability as analyzed in Section 3.5.

Appendix C. Diagnostic Procedures

This appendix summarizes the diagnostic procedures used throughout SORT-AI to identify structural instability, drift accumulation, and fixed-point violations in operator–projection space. All diagnostics are derived directly from the algebraic and kernel constructions introduced in Section 3 and do not rely on architecture-specific assumptions. The procedures are intended to support the use cases discussed in Sections 9–8.

 doi:10.20944/preprints202512.1334.v1

Appendix C.1. Drift Metrics

Structural drift is defined as the cumulative deviation induced by repeated operator application under kernel projection. Given a trajectory $\{x_t\}$ generated by

$$x_{t+1} = \hat{H}\,\Pi_\kappa\left(\sum_i \hat{O}_i x_t\right), \tag{A7}$$

the drift over a horizon T is quantified as

$$\Delta_T = \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|. \tag{A8}$$

Elevated values of Δ_T indicate structural instability or amplification across deep transformation chains. Kernel-weighted variants of Equation A8 are used to emphasize non-local contributions, as discussed in Section 6.3.

Appendix C.2. Collapse and Instability Tests

Operator collapse refers to the effective dominance or elimination of subsets of operators under repeated updates. Collapse is detected by monitoring the operator participation weights

$$w_i(t) = \left\|\Pi_\kappa(\hat{O}_i x_t)\right\|, \tag{A9}$$

and identifying regimes where $w_i(t)$ converges toward zero or saturates disproportionately. Instability is signaled by exponential growth in one or more modes, corresponding to violation of stability conditions such as those in Equation **??**.

Appendix C.3. Fixed-Point Verification

A state x^* is considered a structural fixed point if it satisfies

$$x^* = \hat{H}\,\Pi_\kappa\left(\sum_i \hat{O}_i x^*\right). \tag{A10}$$

Verification proceeds by iterating Equation A7 from perturbed initial conditions and evaluating convergence back to x^*. Failure to return indicates instability or the presence of competing attractors. In SORT-AI, alignment-relevant fixed points are further evaluated for structural robustness by testing sensitivity to kernel scale σ_0 and operator-subset perturbations, consistent with the analyses in Sections 5.2 and 11.1.

Appendix D. Reproducibility and Deterministic Configuration

This appendix documents the reproducibility guarantees and deterministic configuration principles underlying SORT-AI. All structural analyses presented in this work are derived from a fully deterministic operator–projection pipeline inherited from the SORT v5 framework and executed within the MOCK v3 environment [49]. The purpose of this appendix is to clarify the scope of reproducibility claims and to specify which components are fixed, versioned, and verifiable.

Appendix D.1. Deterministic Operator Pipeline

All SORT-AI diagnostics are computed using a deterministic update rule of the form

$$x_{t+1} = \hat{H}\,\Pi_\kappa\left(\sum_i \hat{O}_i x_t\right), \tag{A11}$$

 doi:10.20944/preprints202512.1334.v1

where the operator set $\{\hat{O}_i\}$, the global projector $\hat{H}$, and the projection kernel $\kappa(k)$ are fixed and version-controlled. No stochastic sampling, randomized initialization, or architecture-dependent heuristics enter the operator evolution itself. Given identical initial states and configuration parameters, the resulting trajectories are exactly reproducible.

Appendix D.2. Configuration Parameters and Version Control

All configuration parameters relevant to SORT-AI are specified explicitly and archived, including:

- operator definitions and structure coefficients,
- kernel functional form and correlation scale σ_0,
- numerical tolerances for invariant checks,
- iteration horizons and convergence thresholds.

These parameters are inherited from the SORT v5 reference configuration and are not tuned at the AI-module level. Versioning ensures that all results can be traced unambiguously to a specific algebraic and numerical setup.

Appendix D.3. Global Hash and Archive Integrity

Reproducibility of the complete computational environment is guaranteed through a single global SHA–256 hash associated with the archived MOCK v3 configuration. This hash uniquely identifies the full set of source files, configuration artifacts, and deterministic outputs used in this work. Because all diagnostics are derived from the same fixed pipeline, no file-level hashes are required for independent verification.

Appendix D.4. Scope and Limits of Reproducibility

The reproducibility guarantees provided here apply to the structural diagnostics and mathematical analyses performed within SORT-AI. They do not extend to empirical model behaviors, training outcomes, or evaluation results of specific AI systems, which depend on external data, architectures, and optimization processes. SORT-AI therefore offers reproducible structural analysis rather than reproducible behavioral prediction, consistent with the non-predictive scope discussed in Section 11.

References

1. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv:1906.01820*. arXiv:1906.01820
2. Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv:2401.05566*. arXiv:2401.05566
3. Anthropic (2024). Simple Probes Can Catch Sleeper Agents. Anthropic Alignment Note. anthropic.com/research/probes-catch-sleeper-agents
4. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565*. arXiv:1606.06565
5. Ngo, R., Chan, L., & Mindermann, S. (2022). The Alignment Problem from a Deep Learning Perspective. *arXiv:2209.00626*. arXiv:2209.00626
6. Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk? *arXiv:2206.13353*. arXiv:2206.13353
7. Krakovna, V., et al. (2020). Specification Gaming: The Flip Side of AI Ingenuity. DeepMind Blog. deepmind.com/blog/specification-gaming
8. Olsson, C., et al. (2022). In-context Learning and Induction Heads. *Transformer Circuits Thread*. transformer-circuits.pub
9. Elhage, N., et al. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. transformer-circuits.pub
10. Elhage, N., et al. (2022). Toy Models of Superposition. *Transformer Circuits Thread*. transformer-circuits.pub
11. Olah, C., et al. (2020). Zoom In: An Introduction to Circuits. *Distill*. DOI:10.23915/distill.00024.001
12. Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. *NeurIPS 2023*. arXiv:2304.14997

 doi:10.20944/preprints202512.1334.v1

13. Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress Measures for Grokking via Mechanistic Interpretability. *ICLR 2023*. arXiv:2301.05217
14. Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety – A Review. *arXiv:2404.14082*. arXiv:2404.14082
15. Burns, C., et al. (2023). Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision. *arXiv:2312.09390*. arXiv:2312.09390
16. Bowman, S. R., et al. (2022). Measuring Progress on Scalable Oversight for Large Language Models. *arXiv:2211.03540*. arXiv:2211.03540
17. Leike, J., et al. (2018). Scalable Agent Alignment via Reward Modeling: A Research Direction. *arXiv:1811.07871*. arXiv:1811.07871
18. Irving, G., Christiano, P., & Amodei, D. (2018). AI Safety via Debate. *arXiv:1805.00899*. arXiv:1805.00899
19. Christiano, P., et al. (2018). Supervising Strong Learners by Amplifying Weak Experts. *arXiv:1810.08575*. arXiv:1810.08575
20. Power, A., et al. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *arXiv:2201.02177*. arXiv:2201.02177
21. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. *NeurIPS 2022*. arXiv:2205.10343
22. Rubin, N., Seroussi, I., & Ringel, Z. (2023). Grokking as a First Order Phase Transition in Two Layer Networks. *ICLR 2024*. arXiv:2310.03789
23. Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining Grokking Through Circuit Efficiency. *arXiv:2309.02390*. arXiv:2309.02390
24. Wei, J., et al. (2022). Emergent Abilities of Large Language Models. *TMLR 2022*. arXiv:2206.07682
25. Ganguli, D., et al. (2022). Predictability and Surprise in Large Generative Models. *FAccT 2022*. arXiv:2202.07785
26. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *NeurIPS 2023*. arXiv:2304.15004
27. Shevlane, T., et al. (2023). Model Evaluation for Extreme Risks. *arXiv:2305.15324*. arXiv:2305.15324
28. Phuong, M., et al. (2024). Evaluating Frontier Models for Dangerous Capabilities. *arXiv:2403.13793*. arXiv:2403.13793
29. METR (2024). Autonomy Evaluation Resources. metr.org
30. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828. DOI:10.1109/TPAMI.2013.50
31. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press. ISBN 978-0-262-17005-5.
32. Koh, P. W., et al. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. *ICML 2021*. arXiv:2012.07421
33. Reed, M., & Simon, B. (1980). *Methods of Modern Mathematical Physics I: Functional Analysis* (Revised ed.). Academic Press. ISBN 978-0-12-585050-6.
34. Kato, T. (1995). *Perturbation Theory for Linear Operators* (Reprint of 1980 ed.). Springer. ISBN 978-3-540-58661-6.
35. Halmos, P. R. (1982). *A Hilbert Space Problem Book* (2nd ed.). Springer. ISBN 978-0-387-90685-0.
36. Bhatia, R. (1997). *Matrix Analysis*. Springer. ISBN 978-0-387-94846-1.
37. Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *NeurIPS 2018*. arXiv:1806.07572
38. Arora, S., Du, S. S., Hu, W., Li, Z., & Wang, R. (2019). Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *ICML 2019*. arXiv:1901.08584
39. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep Double Descent: Where Bigger Models and More Data Can Hurt. *J. Stat. Mech.* **2021**, 124003. arXiv:1912.02292
40. Vaswani, A., et al. (2017). Attention Is All You Need. *NeurIPS 2017*. arXiv:1706.03762
41. Brown, T., et al. (2020). Language Models are Few-Shot Learners. *NeurIPS 2020*. arXiv:2005.14165
42. Anthropic (2024). The Claude Model Card and Evaluations. anthropic.com
43. Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *NeurIPS 2022*. arXiv:2203.02155
44. Bai, Y., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862*. arXiv:2204.05862
45. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*. arXiv:2305.18290

 doi:10.20944/preprints202512.1334.v1

46. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ICLR 2015*. arXiv:1412.6572
47. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018*. arXiv:1706.06083
48. Hendrycks, D., et al. (2021). Unsolved Problems in ML Safety. *arXiv:2109.13916*. arXiv:2109.13916
49. Wegener, G. H. (2025). Supra-Omega Resonance Theory: A Nonlocal Projection Framework for Cosmological Structure Formation. *Whitepaper v5*. Zenodo. DOI:10.5281/zenodo.17787754
50. Ji, J., et al. (2023). AI Alignment: A Comprehensive Survey. *arXiv:2310.19852*. arXiv:2310.19852
51. Perez, E., et al. (2022). Red Teaming Language Models with Language Models. *arXiv:2202.03286*. arXiv:2202.03286
52. Greenblatt, R., et al. (2023). AI Control: Improving Safety Despite Intentional Subversion. *arXiv:2312.06942*. arXiv:2312.06942
53. Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and Characterizing Reward Hacking. *NeurIPS 2022*. arXiv:2209.13085
54. Langosco, L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022). Goal Misgeneralization in Deep Reinforcement Learning. *ICML 2022*. arXiv:2105.14111
55. Pan, A., Bhatia, K., & Steinhardt, J. (2022). The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. *ICLR 2022*. arXiv:2201.03544
56. Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv:2310.01405*. arXiv:2310.01405
57. Pan, A., et al. (2023). Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. *ICML 2023*. arXiv:2304.03279
58. Perez, E., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. *ACL 2023*. arXiv:2212.09251
59. Clymer, J., et al. (2025). Safety Pretraining: Toward the Next Generation of Safe AI. *https://doi.org/10.48550/arXiv.2504.16980*.
60. Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv:2309.08600*. arXiv:2309.08600
61. Bricken, T., et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. transformer-circuits.pub
62. Templeton, A., et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic*. anthropic.com