The Structural Coherence Bound

A Formal Derivation of MAI-1 Invariant 4:
Dirichlet Energy Band Specification,
Representational Health Monitoring, and
Inference-Time Geometric Certification

Clause AI-7

Fields

Specification — Version 1.0

February 2026

Abstract

The deployment of foundation models in regulated industries—financial services, healthcare, defense, and critical infrastructure—demands continuous verification not only of a model's output behavior but of the geometric health of its internal representations. A model may maintain stable output entropy, bounded gradient norms, low distribution drift, and clean hardware attestation while its representational geometry silently degrades: internal features collapsing to a low-dimensional subspace incapable of distinguishing meaningfully different inputs, or fragmenting into chaotically sensitive manifolds exploitable by adversarial perturbation. Neither failure mode is detectable by the four existing MAI-1 invariants. The Model Attestation Interface (MAI-1; Clause AI-5, Auburn Patent Family; Fields, 2026) defines five mandatory health invariants, of which Invariant 4—Structural Coherence—requires that the Dirichlet energy of the model's internal representations remain within a certified coherence band $[E^D_{\min}, E^D_{\max}]$ at all attested measurement points. MAI-1 defines the `coherence-energy` field, its governance rationale, and its breach semantics, but does not derive the bound, specify the measurement methodology, or define the calibration procedure for establishing the energy band.

This document fills that gap. It provides the complete mathematical foundation for MAI-1's structural coherence invariant: the spectral graph-theoretic definition of Dirichlet energy and its adaptation to transformer architectures via the attention-as-dynamic-graph formulation; the formal characterization of the two failure modes that coherence monitoring detects—representational collapse (energy below the certified floor, linked to oversmoothing, neural collapse, dimensional collapse, token uniformity, and loss of plasticity) and representational fragmentation (energy above the certified ceiling, linked to oversquashing, over-sharpening, chaotic dynamics, and adversarial vulnerability); the effective rank complementary diagnostic that captures representational degradation in subspaces orthogonal to graph structure; the layer-wise measurement methodology including attention-weighted energy computation, head aggregation, and the DE Ratio layer-wise diagnostic; the efficient inference-time computation architecture—Top-K sparsification, stochastic trace estimation, randomized SVD, and Frobenius proxy metrics—that makes coherence monitoring feasible at scale within a $< 2\%$ latency budget; the calibration procedure for establishing the certified coherence band from validation data with conformal coverage guarantees; and the compliance state machine with deterministic transition rules.

The result is a governance-grade specification that transforms the qualitative concern of "representational health" into an auditable, insurable, and enforceable property of the deployed model. With Clause AI-7 enforced alongside AI-8 (Entropy Collapse Constraint), AI-2 (Gradient Starvation Envelope), AI-6 (Distribution Drift Bound), and AI-4 (SRAM Thermal Integrity Bound), all five mandatory MAI-1 invariants are formally derived with full calibration methodology, completing the analytical core of the Auburn Governance Stack's Layer 2 invariant suite.

Implementation Notice

Clause AI-7 Is Not a Standalone Specification

DO NOT IMPLEMENT THIS DOCUMENT IN ISOLATION

This document specifies **one invariant** within a 45-document governance infrastructure. Implementing Clause AI-7 without the upstream specifications it depends on and the downstream documents it feeds into will produce a monitoring system that is **structurally incomplete**, **cryptographically unbound**, and **non-conformant** with the Auburn Governance Stack.

Partial implementation creates a false sense of governance assurance. A Dirichlet energy monitor without the MAI-1 attestation binding is a research tool, not a governance instrument. The distinction matters.

Upstream Dependencies (Required by AI-7)

Clause AI-7 assumes the existence and conformance of the following specifications. Without these, the invariant's governance guarantees do not hold.

Document	Title	What AI-7 Requires From It
MSAF	Model State Attestation Framework	Three-tier attestation architecture that defines *where* invariant measurements are anchored (hardware root of trust, model state, provenance). Without MSAF, AI-7 measurements float in air—attested by nothing.
MAI-1 (AI-5)	Model Attestation Interface	Invariant 4 definition, Layer 2 payload schema, attestation token format, CoRIM binding mechanism. Without MAI-1, AI-7 measurements have no standardized delivery format and no cryptographic seal.
AGS-1	Governance Stack Master Architecture	Layer definitions, composition waist, dependency structure. Without AGS-1, AI-7 has no architectural context—it is a leaf without a tree.

Downstream Consumers (Fed by AI-7)

The following documents consume AI-7's outputs. Implementing AI-7 without awareness of these consumers produces measurements that nothing acts on.

Document	Title	What It Requires From AI-7
CTS-1	Conformance Test Suite	The coherence band is a testable pass/fail criterion. CTS-1 defines the binary conformance assertions (AS-SM-L2-20 and related) that make AI-7 compliance *checkable by strangers*. Without CTS-1, AI-7 compliance is self-reported and unverifiable.
MAI-1 Layer 2 Payload	Attestation Token	The `coherence-energy`, `coherence-rank`, `coherence-fiedler`, and `coherence-state` fields must be carried in the standardized MAI-1 token. Without the payload integration, AI-7 measurements are local telemetry, not governance evidence.
Document 31	Adversarial Robustness Testing	Coherence energy changes under adversarial attack serve as detection signals. Document 31 specifies how AI-7 measurements are consumed as inputs to adversarial robustness evaluation.
Sector Profiles	EU AI Act, FDA, Financial Services, etc.	The Application Layer profiles map AI-7 compliance to specific regulatory requirements. Without these profiles, AI-7 conformance cannot be translated into regulatory language.

Cross-Invariant Dependencies

AI-7 is the fifth of five mandatory invariants. It does not function as intended without the other four.

Invariant	Clause	Integration Dependency
AI-8	Entropy Collapse Constraint	Composite failure signatures (Section 12.3) require correlated analysis of entropy and energy. AI-7 without AI-8 cannot distinguish representational collapse from output miscalibration.
AI-2	Gradient Starvation Envelope	Cross-invariant triggering (Section 11.3.3) uses gradient anomalies to trigger enhanced coherence monitoring during fine-tuning.
AI-4	SRAM Thermal Integrity Bound	Hardware-caused representational corruption (Signature 3, Section 12.3.3) requires AI-4 co-analysis to distinguish thermal bit errors from model-caused pathology.
AI-6	Distribution Drift Bound	Complementary coverage (Section 12.2.1): distributional drift detects statistical shift while AI-7 detects geometric degradation. Neither subsumes the other.

The Composition Principle

The Auburn Governance Stack is an **infrastructure specification**, not a paper series. Each document is a component in a composition architecture. Clause AI-7 feeds into the MAI-1/AGS-1 composition waist. Systems that do not expose an MAI-1 compliant endpoint are outside the scope of all downstream Auburn governance clauses.

For academic researchers: This document is freely available for non-commercial study, citation, and analysis under CC BY-NC-ND 4.0. The geometric framework, measurement methodology, and empirical grounding stand on their own as contributions to the representational health literature.

UncleBroFields@proton.me | **fieldsryanchristopher@gmail.com**

Contents

1 Introduction: The Representational Health Problem

1.1 The Operational Question

Every foundation model deployed in a regulated environment must answer a question that no current monitoring infrastructure can address:

> *Is the geometric structure of this model's internal representations healthy right now—or has it silently collapsed into a low-dimensional subspace, fragmented into adversarially exploitable manifolds, or entered a chaotic regime from which generalization is impossible?*

This is not a question about outputs. A model can produce fluent, confident, benchmark-passing text while its internal representation geometry has degraded to a point where catastrophic failure is one distributional shift away. Output-level metrics—perplexity, accuracy, F1 score, and even the distribution drift monitored by Clause AI-6—are necessary but insufficient. They measure the *symptoms* of model health. Structural coherence measures the *anatomy*.

The distinction is analogous to cardiovascular medicine: blood pressure and heart rate (output metrics) can appear normal while arterial plaque accumulates silently (structural degradation). By the time output metrics detect the problem, the damage may be irreversible. The governance question is whether the model's internal geometry—the manifold on which it represents, transforms, and reasons about data—remains within the operational regime established during validation.

1.2 The Governance Evidence Vacuum for Structural Health

Current AI governance frameworks universally require "robustness" and "consistent performance throughout the lifecycle" (EU AI Act, Article 15), "ongoing monitoring confirming models perform as intended" (SR 11-7), and "quantitative tools and methodologies for monitoring" (NIST AI RMF, MEASURE function). Yet no deployed monitoring system measures the geometric health of a model's internal representations. The entire industry monitors outputs and hopes the internals are fine.

This creates a specific and dangerous evidence vacuum:

- **Regulators** cannot verify that a model's internal structure is stable because no metric exists to measure it. Article 15(2) of the EU AI Act calls for "benchmarks and measurement methodologies" for accuracy and robustness—but no standardized methodology addresses representational geometry.
- **Insurers** cannot price the risk of representational degradation because no quantifiable property captures it. The parametric insurance structures emerging from Munich Re and Armilla AI require measurable triggers—not subjective assessments of "model health."
- **Model vendors** cannot demonstrate that their models maintain structural integrity under deployment stress because no attestation artifact carries this evidence. The MAI-1 Layer 2 payload defines the `coherence-energy` field precisely to fill this gap—but without the derivation and calibration methodology specified in this document, the field cannot be populated.
- **Procurement officers** cannot specify structural health requirements in RFPs because no specification exists to reference. The self-authorizing document principle of the Auburn Governance Stack demands that this specification be complete enough to drop into a procurement requirement without explanation.

1.3 What This Document Provides

Clause AI-7 resolves the representational health evidence vacuum by specifying the complete technical infrastructure for continuous structural coherence monitoring within the Auburn Governance Stack's MAI-1 attestation architecture. The specification provides:

1. **A geometric health metric** (Section 3): The Dirichlet energy functional adapted from spectral graph theory to transformer architectures via the attention-as-dynamic-graph formulation, providing a unified, architecture-aware measure of representational smoothness and diversity.

2. **A formal characterization of two failure modes** (Sections 4–5): Representational collapse (energy below the certified floor) and representational fragmentation (energy above the certified ceiling), grounded in the established literatures on oversmoothing, neural collapse, dimensional collapse, oversquashing, and adversarial robustness.

3. **A mandatory invariant** (Section 6): $E^D_{\min}(\ell) \leq E_D(X^{(\ell)}) \leq E^D_{\max}(\ell)$ at every attested measurement point, with the coherence band calibrated from the model's own empirical distribution using conformal methods.

4. **A transformer-specific measurement methodology** (Section 7): The exact computation of attention-weighted Dirichlet energy, head aggregation strategy, layer selection protocol, the DE Ratio layer-wise diagnostic, and the effective rank complementary metric.

5. **An efficient inference-time computation architecture** (Section 8): The calibration procedure for establishing the certified coherence band $[E^D_{\min}, E^D_{\max}]$ from validation data.

6. **A computational tractability proof** (Section 9): Top-K sparsification, stochastic trace estimation (Hutchinson's method), randomized SVD, and Frobenius proxy metrics demonstrating $< 2\%$ latency overhead under production conditions.

7. **A six-framework regulatory mapping** (Section 10): Citation-level mapping to EU AI Act, SR 11-7, FDA PCCP, NIST AI RMF, Solvency II, and emerging frontier safety frameworks.

8. **A production-ready implementation architecture** (Section 11): The Coherence Monitor with three subsystems (Energy Estimator, Rank Classifier, State Controller), automated intervention mechanisms, and a self-authorizing MAI-1 Layer 2 attestation payload.

9. **A cross-invariant integration analysis** (Section 12): How AI-7 completes the five-invariant set and interacts with AI-8, AI-2, AI-6, and AI-4 to provide comprehensive Layer 2 health monitoring.

10. **An insurability architecture** (Section 13): The certified coherence band as an insurable property with parametric trigger structure.

1.4 Dependency Structure

Clause AI-7 occupies a specific position in the Auburn Governance Stack's dependency graph:

Dependency Declaration

Feeds into:

- MAI-1 §7.4 (Invariant 4: Structural Coherence), providing the derivation and calibration methodology for the `coherence-energy` field.
- MAI-1 Layer 2 payload, populating the `coherence-energy` and `thresholds.coherence-energy-max` fields.
- CTS-1 (Conformance Test Suite), providing testable pass/fail criteria for coherence band compliance.
- Document 31 (Adversarial Robustness Testing), providing the geometric basis for adversarial detection via energy band violation.

Requires:

- MSAF (Three-Tier Architecture): The model health invariant architecture within which AI-7 operates.
- MAI-1 (Clause AI-5): The interface specification defining the `coherence-energy` field, its data type, and its governance semantics.
- AI-6 (Distribution Drift Bound): The conformal calibration methodology that AI-7 adapts for energy band construction.
- AI-8 (Entropy Collapse Constraint): The entropy floor invariant whose interaction with structural coherence is formally characterized in Section 12.

1.5 Relationship to Completed Invariants

Clause AI-7 is the fifth and final mandatory invariant in the MAI-1 specification. The four completed invariants and their relationship to structural coherence are:

Invariant	What It Monitors	What It Misses	AI-7 Fills the Gap
AI-8: Entropy Collapse	Output diversity (token distribution entropy)	A model can maintain output entropy while internal representations collapse to a subspace	AI-7 detects the internal collapse that precedes entropy floor violation
AI-2: Gradient Starvation	Training stability (gradient norm bounds)	Gradient norms can stabilize while representational geometry fragments under adversarial pressure	AI-7 detects geometric fragmentation independent of gradient behavior
AI-6: Distribution Drift	Behavioral consistency (KL divergence from baseline)	Drift monitoring compares output distributions; a model can maintain low output drift while internal geometry degrades	AI-7 monitors the representational structure that *produces* the outputs

Invariant	What It Monitors	What It Misses	AI-7 Fills the Gap
AI-4: SRAM Thermal	Hardware integrity (thermal side-channel bounds)	Hardware can be perfectly healthy while the model running on it has collapsed representations	AI-7 monitors model-level geometry independent of hardware state

Table 1: Relationship of AI-7 to the four completed MAI-1 invariants. Each existing invariant monitors a necessary dimension of model health that is insufficient to detect structural coherence degradation.

The critical insight is that a model can simultaneously satisfy all four existing invariants—stable entropy, bounded gradients, low output drift, clean hardware—while its internal representational geometry has degraded to a pathological state. This occurs because the four invariants monitor *different projections* of the model's state: output statistics (AI-8, AI-6), optimization dynamics (AI-2), and hardware integrity (AI-4). None monitors the *geometric structure of the latent space itself*. Clause AI-7 fills this gap.

1.6 Document Structure

The remainder of this document is organized as follows:

- **Section 2** presents the phenomenology of structural degradation—documented cases from production and research where representational geometry fails, organized by failure mechanism with explicit Clause AI-7 interpretations.
- **Section 3** develops the geometric foundations: the Dirichlet energy functional, spectral graph theory, the attention-as-dynamic-graph bridge to transformers, and the formal definition of structural coherence.
- **Section 4** formally characterizes the lower bound failure mode—representational collapse—drawing on neural collapse, dimensional collapse, transformer rank collapse, oversmoothing, and loss of plasticity.
- **Section 5** formally characterizes the upper bound failure mode—representational fragmentation—drawing on oversquashing, curvature analysis, chaotic dynamics, and adversarial vulnerability.
- **Section 6** presents the mathematical formalization of Clause AI-7: the structural coherence invariant, the coherence band, complementary diagnostics, the compliance state machine, and the audit-ready parameter table.
- **Section 7** specifies the transformer-specific measurement methodology: attention-weighted energy computation, head aggregation, layer selection, and the effective rank computation.
- **Section 8** defines the calibration procedure for establishing the certified coherence band from validation data.
- **Section 9** proves computational tractability: the efficient computation architecture that achieves $< 2\%$ latency overhead.
- **Section 10** maps the specification to six regulatory frameworks at citation level.

- **Section 11** specifies the production implementation architecture: the Coherence Monitor and its integration with the MAI-1 Layer 2 attestation payload.
- **Section 12** analyzes cross-invariant integration: how AI-7 completes the five-invariant set and the composite failure signatures that emerge from joint monitoring.
- **Section 13** develops the insurability architecture: the certified coherence band as a parametric trigger for AI insurance products.
- **Section 14** presents limitations, future work, and the conclusion.

Honest Framing

Clause AI-7 provides *geometric health monitoring*—continuous verification that the model's internal representational structure remains within the operational regime established during validation. It does not guarantee behavioral safety, prevent all adversarial attacks, or certify that the model's outputs are correct. The coherence band certifies that the model's representational geometry has not degraded to a pathological state. A model within the certified band may still produce incorrect outputs; a model outside the band is certifiably operating in a regime where its representational structure has failed. This is the distinction between health monitoring and behavioral guarantee, and it is maintained throughout this specification.

2 The Phenomenology of Structural Degradation: Evidence from Production and Research

The structural coherence invariant is not a theoretical exercise. Each failure mode it monitors has been documented in production systems, empirical research, or both. This section catalogs the known pathologies of representational geometry, organized by mechanism, with explicit Clause AI-7 interpretations that connect each phenomenon to the certified coherence band.

2.1 Token Uniformity in Deep Transformers

2.1.1 The Mechanism

Dong, Cordonnier, and Loukas (2021) proved that pure self-attention—without residual connections or feed-forward networks—converges to a rank-1 output matrix at a *doubly exponential* rate in depth. In this terminal state, all token representations are identical: $x_i = x_j$ for all tokens i, j. The Dirichlet energy of such a representation is exactly zero.

This result is far more severe than the exponential oversmoothing observed in graph neural networks. While architectural components—residual connections, LayerNorm, and feed-forward layers—are designed to counteract this convergence, they do not eliminate it. Empirical analysis of BERT, GPT-2, and XLNet confirms that removing skip connections causes rapid rank collapse across all architectures. Even with skip connections intact, the cosine similarity between token representations in the final layers of production vision transformers (DeiT) approaches $\sim$0.9, indicating substantial—though not complete—convergence toward uniformity.

2.1.2 The Architectural Countermeasures and Their Limits

Residual connections are the primary defense against token uniformity. By adding the input to the output at each layer ($X^{(\ell+1)} = X^{(\ell)} + F(X^{(\ell)})$), they preserve a "residual stream" that maintains token identity across depth. However, this defense is not absolute:

- LayerNorm does not prevent rank collapse. As Dong et al. demonstrated, right-multiplication by a matrix (which LayerNorm implements in the feature dimension) cannot increase the rank of the representation matrix. LayerNorm rescales but does not restore lost dimensions.
- The balance between the attention mechanism (which mixes tokens and reduces energy) and the residual stream (which preserves token identity and maintains energy) is architecture-dependent and can shift during fine-tuning. A model validated with healthy energy balance can lose that balance after domain-specific fine-tuning.
- Pre-LayerNorm configurations exhibit power-law (rather than exponential) energy scaling, avoiding catastrophic collapse but introducing a "curse of depth" where representations evolve minimally in deeper layers. Post-LayerNorm configurations avoid this stagnation but are more susceptible to rapid oversmoothing.

2.1.3 The Clause AI-7 Interpretation

Token uniformity is the transformer-specific manifestation of the *collapse floor* violation ($E_D < E_{\min}^D$). When the attention mechanism overwhelms the preserving force of the residual stream, internal representations converge toward a centroid, and the model loses the capacity to distinguish between semantically distinct inputs. AI-7 detects this by monitoring the Dirichlet energy computed over the attention-weighted token graph. A sustained decline in energy toward the floor—particularly in middle layers where collapse typically initiates—triggers a compliance state transition before the collapse propagates to output-affecting layers.

2.2 Neural Collapse in Language Models

2.2.1 The Phenomenon

Neural collapse, identified by Papyan, Han, and Donoho (2020), describes the geometric structure that emerges in the terminal phase of training in classification networks. The phenomenon consists of four interconnected properties: within-class variability vanishes (NC1), class means converge to a Simplex Equiangular Tight Frame (NC2), classifier weights align with class means (NC3), and classification reduces to nearest-class-center assignment (NC4). In the classification setting, neural collapse at the *last layer* is optimal—it maximizes inter-class separation and represents the theoretically ideal geometry for the task.

The pathology arises when collapse occurs *prematurely*—in intermediate layers—or *excessively*—destroying the hierarchical feature representations required for transfer learning, out-of-distribution generalization, and complex reasoning.

2.2.2 Neural Collapse in the LLM Regime

Wu and Papyan (2024) investigated neural collapse in causal language models, where none of the standard conditions hold: the vocabulary ("class count") vastly exceeds the embedding dimension, training runs for far fewer epochs than in the classification regime, and the loss function operates over sequences rather than individual samples. Despite these differences, they observed *partial* neural collapse emergence—a critical finding that establishes the phenomenon as relevant to the LLM paradigm.

In language models, excessive collapse manifests as *linguistic collapse*: the embedding space for long-tail concepts—rare words, specialized terminology, nuanced semantic distinctions—degenerates. The model retains the capacity to produce common tokens fluently while losing the representational resolution required for specialized knowledge. This degradation is invisible to output-level metrics evaluated on standard benchmarks, which are dominated by high-frequency vocabulary.

2.2.3 The Clause AI-7 Interpretation

Neural collapse in intermediate layers produces a sharp drop in Dirichlet energy: the representation manifold projects onto a low-dimensional subspace, and the pairwise distances between token representations decrease. AI-7 monitors for this energy signature. The effective rank complementary diagnostic (Section 6) provides additional resolution: a layer exhibiting low Dirichlet energy *and* low effective rank is in a collapse state, while low energy with maintained rank may indicate healthy convergence in the final layer. The distinction between healthy terminal-phase collapse and pathological intermediate collapse is encoded in the *layer-specific* nature of the coherence band—$E^D_{\min}(\ell)$ is calibrated independently for each monitored layer.

2.3 Dimensional Collapse in Foundation Models

2.3.1 The Mechanism

Jing, Vincent, LeCun, and Tian (2022) proved that contrastive and non-contrastive self-supervised learning methods can drive the learned representations to span a lower-dimensional subspace than the available ambient dimension. A model with an embedding dimension of 4096 may effectively utilize only 50 dimensions—a "capacity famine" where billions of parameters contribute nothing to representational diversity.

Two root causes drive dimensional collapse: strong data augmentation makes the covariance matrix of the augmented views low-rank, and implicit regularization in over-parameterized networks drives weights toward low-rank solutions. The result is that the singular values of the representation covariance matrix decay rapidly, with a small number of dominant directions capturing nearly all variance.

2.3.2 The Tunnel Effect

Masarczyk et al. (2023) identified a related phenomenon in supervised deep networks: sufficiently deep networks split into an "extractor" (early layers that create linearly separable representations) and a "tunnel" (later layers that compress representations to the minimum numerical rank needed for the task—approximately equal to the number of classes). The tunnel *degrades* out-of-distribution generalization and transfer learning performance because it discards the rich, high-dimensional features learned by the extractor.

This is particularly dangerous for foundation models intended for multi-task deployment: a model fine-tuned on a narrow task develops a tunnel that destroys the general-purpose representational capacity the model was originally validated on. The fine-tuned model passes task-specific benchmarks while having fundamentally different—and degraded—internal geometry.

2.3.3 The Clause AI-7 Interpretation

Dimensional collapse and the tunnel effect both manifest as a decline in effective rank without necessarily producing a proportional decline in Dirichlet energy. This is the critical insight from Zhang, Wei, Xu, and Liu (2025), who proved that rank collapse is *strictly more general* than energy collapse: representations can lose dimensionality (rank drops) while remaining diverse along the surviving dimensions (energy stays moderate). AI-7 addresses this by mandating *both* the Dirichlet energy band and the effective rank floor as complementary diagnostics. A model that maintains energy within $[E^D_{\min}, E^D_{\max}]$ but drops below the effective rank threshold is flagged as exhibiting dimensional collapse—a state invisible to energy monitoring alone.

2.4 Oversmoothing in Deep Architectures

2.4.1 The Graph Neural Network Evidence

The oversmoothing pathology was first documented in graph neural networks by Li, Han, and Wu (2018), who demonstrated that graph convolution is a special form of Laplacian smoothing. As GNNs become deeper, the repeated application of this smoothing operator forces node representations to converge to a non-informative subspace. Performance on standard benchmarks (e.g., Cora) degrades dramatically beyond approximately four layers.

Oono and Suzuki (2020) provided the first rigorous proof that this convergence is *exponential*: $\|X^{(\ell)} - X^*\| \leq C_1 \exp(-C_2 \ell)$, where C_2 depends on the spectral gap of the graph Laplacian and the spectral norm of the weight matrices. The Dirichlet energy decays at a rate governed by $e^{-2\lambda_1 t}$ in the continuous diffusion limit, where λ_1 is the first nonzero Laplacian eigenvalue.

2.4.2 The Transfer to Transformers

The oversmoothing analogy transfers to transformers through the attention-as-graph formulation. Self-attention layers act as adaptive low-pass filters on the token representation graph: they mix information across tokens, weighted by the attention distribution. When the mixing is too aggressive—high attention entropy distributing weight broadly across all tokens—the token representations converge toward a common centroid.

Guo et al. (2023, NeuTRENO) formalized this connection by proving that self-attention layers minimize a nonlocal functional that promotes smoothness—directly analogous to the graph Dirichlet energy functional. Their empirical measurements on vision transformers (DeiT) showed cosine similarity between token representations approaching $\sim$0.9 in final layers, confirming that oversmoothing is not merely a theoretical concern but a measurable production phenomenon.

2.4.3 The Clause AI-7 Interpretation

Oversmoothing is the *canonical* collapse-floor pathology. In the Dirichlet energy framework, oversmoothing corresponds to $E_D \to 0$: connected nodes (tokens with high mutual attention weight) converge to identical representations, and the energy functional—which measures pairwise squared differences weighted by edge strength—approaches zero. AI-7 sets $E^D_{\min}$ to guard against this state. The layer-wise DE Ratio diagnostic ($R_{DE} = E_D(\text{Output})/E_D(\text{Input})$ for each layer) provides granular visibility: a sustained $R_{DE} < 1.0$ across consecutive layers indicates progressive smoothing. A product of DE Ratios across depth that approaches zero predicts imminent collapse.

2.5 Loss of Plasticity and Representational Rigidity

2.5.1 The Phenomenon

Dohare, Hernandez-Garcia, Lan, Rahman, Mahmood, and Sutton (2024) demonstrated in *Nature* that loss of plasticity—the progressive inability of a neural network to learn from new data—is pervasive in continual learning settings. They identified three correlates of plasticity loss: increasing weight magnitude, *decreasing effective rank* of activation matrices, and increasing fraction of dead units.

Lyle, Zheng, Nikishin, Pires, Pascanu, and Dabney (2023) provided a complementary analysis showing that no single measured quantity—weight norm, dead units, or feature rank—reliably predicts plasticity loss across all settings. However, the collective pattern is consistent: representational degradation accompanies the loss of learning capacity.

2.5.2 Two Types of Plasticity Failure

The plasticity loss literature distinguishes two dynamical regimes:

- **Type-1 (Contractive/Collapse):** The network's dynamics become overly contractive. Representations collapse into a low-dimensional attractor. Gradients vanish in the collapsed directions, and the model becomes rigid—incapable of adapting to new distributional conditions. This corresponds to the *energy floor* violation: Dirichlet energy trends toward $E^D_{\min}$ as representational diversity is destroyed.
- **Type-2 (Expansive/Chaotic):** The network's dynamics become expansive. Infinitesimally close inputs diverge exponentially in the representation space. The model becomes brittle—memorizing training data but failing to generalize, as the smooth interpolation properties of the manifold are destroyed. This corresponds to the *energy ceiling* violation: Dirichlet energy trends toward $E^D_{\max}$ as representational stability breaks down.

2.5.3 The Clause AI-7 Interpretation

Loss of plasticity is a *dynamic* pathology: it develops over time as the model processes data in deployment or undergoes continued training. AI-7's continuous monitoring detects both types as they develop. A Dirichlet energy trace trending toward the floor is a leading indicator of Type-1 plasticity loss, enabling early intervention (e.g., targeted weight reinitialization or regularization adjustment) before the model becomes permanently rigid. A trace trending toward the ceiling signals Type-2 instability. The effective rank diagnostic is particularly valuable here: Dohare et al. observed that effective rank on Online Permuted MNIST steadily decreases after each task under standard backpropagation (reaching ∼5 from an initial ∼15–20), while interventions that maintain plasticity also maintain stable effective rank near initial values.

2.6 Fine-Tuning-Induced Geometric Degradation

2.6.1 The Mechanism

Foundation models are validated on broad distributions but deployed after domain-specific fine-tuning. This fine-tuning can fundamentally alter the model's representational geometry in ways that are invisible to task-specific benchmarks:

- **Catastrophic forgetting:** The model's representations for previously learned concepts are overwritten by fine-tuning gradients. The effective rank of intermediate layers may drop as the model specializes, creating the tunnel effect described in Section 2.3.
- **Representational anisotropy:** Godey et al. (2024) demonstrated that anisotropy—where representations cluster in a narrow cone rather than occupying the full ambient space—is inherent to self-attention in transformers, regardless of training objective. Fine-tuning on narrow distributions can amplify this effect, concentrating representational diversity in a small number of directions.
- **Adversarial vulnerability shift:** Su, Zhang, Tsilivis, and Kempe (2023) showed that the Simplex ETF structure produced by neural collapse disappears under small adversarial perturbations. Fine-tuned models that achieve clean neural collapse geometry at the last layer may be *more* vulnerable to adversarial attack than the original foundation model, because the collapse concentrates representations in a low-dimensional subspace that is easier to perturb.

2.6.2 The Baseline Versioning Challenge

Fine-tuning creates a versioning challenge for the coherence band: the energy band calibrated for the original foundation model may not apply to the fine-tuned variant. AI-7 addresses this through the same baseline versioning mechanism established in Clause AI-6 (the Reilly Sentinel

Protocol): each fine-tuned model version **SHALL** undergo coherence band recalibration, with the recalibrated band cryptographically bound to the model version via the MAI-1 CoRIM artifact.

2.6.3 The Clause AI-7 Interpretation

Fine-tuning is the most common pathway to structural degradation in production. A foundation model that passes all five MAI-1 invariants at the time of initial deployment can fail the structural coherence invariant after fine-tuning—even if it *improves* on task-specific metrics. AI-7 requires recalibration after any model modification, and the coherence band provides a quantitative check on whether fine-tuning has preserved the representational health of the original model or degraded it. This is the "geometric provenance" guarantee: the fine-tuned model's representational structure is certifiably within the operational regime of a validated system.

2.7 Adversarial Perturbation and Spectral Disruption

2.7.1 The Mechanism

Adversarial attacks inject perturbations designed to maximize the change in internal representations—effectively maximizing the gradient of the loss with respect to the input. In the Dirichlet energy framework, this corresponds to injecting high-frequency noise into the token representation graph: the adversarial perturbation creates large pairwise differences between representations that should be similar, spiking the energy toward the ceiling.

Ma et al. (2018) demonstrated that adversarial regions in the input space have *significantly higher* local intrinsic dimensionality than clean data regions, with detection AUC exceeding 0.90 across five different attack strategies. This finding has a direct Dirichlet energy interpretation: high local intrinsic dimensionality corresponds to high-frequency, high-energy representational structure—the model's manifold has become rough and fragmented in the neighborhood of the adversarial example.

2.7.2 The Spectral Signature of Adversarial Attack

Adversarial attacks produce a characteristic spectral signature in the attention-weighted representation graph:

- **Energy spike:** The Dirichlet energy increases sharply as adversarial perturbations create large representation differences between tokens that the attention mechanism connects. The energy may exceed $E^D_{\max}$.
- **High-Frequency Energy Ratio (HFER) increase:** The proportion of total energy concentrated in high-frequency eigenvectors of the attention Laplacian increases. Clean inputs produce representations dominated by low- and mid-frequency components (smooth, semantically coherent structure). Adversarial inputs shift energy toward high-frequency components (noise, fragmentation).
- **Fiedler value drop:** The algebraic connectivity of the attention graph decreases as the adversarial perturbation causes the attention mechanism to fragment—attending to local neighborhoods rather than integrating global context. This disconnection produces isolated subgraphs in the attention pattern, reducing λ_2.

2.7.3 The Clause AI-7 Interpretation

The $E^D_{\max}$ ceiling serves a dual purpose: it guards against both endogenous fragmentation (chaotic dynamics, over-sharpening) and exogenous perturbation (adversarial attack). By enforcing the energy ceiling, AI-7 acts as a geometric defense mechanism that is *attack-agnostic*: it detects the representational consequence of adversarial perturbation regardless of the specific attack method employed. This does not replace dedicated adversarial defenses but provides an independent, orthogonal detection channel rooted in representational geometry rather than input-space heuristics.

2.8 Quantization-Induced Representational Distortion

2.8.1 The Mechanism

Model quantization—reducing weight and activation precision from FP32/FP16 to INT8, INT4, or lower—is standard practice for deployment efficiency. Quantization introduces systematic rounding errors that can alter the representational geometry of the model:

- Weight quantization truncates the least significant bits of weight matrices, potentially reducing their effective rank and altering their spectral properties. If the quantized weight matrices have systematically different spectral norms than the original, the energy dynamics of the network change.
- Activation quantization clips the dynamic range of internal representations, which can either compress the representation manifold (reducing energy toward the floor) or introduce quantization noise (increasing energy toward the ceiling), depending on the clipping strategy and the distribution of activation values.

2.8.2 The Clause AI-7 Interpretation

Quantization is a controlled form of representational perturbation. AI-7 provides a quantitative check on whether quantization has preserved the structural coherence of the original model: the quantized model's Dirichlet energy profile **SHALL** remain within the coherence band calibrated for the full-precision model, or the band **SHALL** be recalibrated for the quantized variant. This transforms quantization validation from "does the quantized model match the original's benchmark scores?" to "does the quantized model maintain the original's representational geometry?"—a strictly stronger test.

2.9 Summary of Documented Structural Degradation Pathologies

Pathology	**Failure Mode**	**Energy Signature**	**Rank Signature**	**AI-7 Detection**
Token uniformity	Collapse	$E_D \to 0$	Rank $\to 1$	Floor violation
Neural collapse (intermediate)	Collapse	$E_D \ll E^D_{\min}$	Rank $\to C$ (classes)	Floor + rank drop
Dimensional collapse	Collapse	May appear normal	Rank $\ll d$	Rank floor violation
Oversmoothing	Collapse	Exponential decay	Progressive decline	DE Ratio < 1 sustained
Tunnel effect	Collapse	Moderate	Rank $\to C$ in deep layers	Layer-specific rank drop

Pathology	Failure Mode	Energy Signature	Rank Signature	AI-7 Detection
Type-1 plasticity loss	Collapse	Trending to floor	Decreasing	Trend detection
Adversarial perturbation	Fragmentation	$E_D \gg E^D_{\max}$	LID increases	Ceiling violation
Over-sharpening	Fragmentation	Energy spike	Moderate–high	Ceiling + HFER spike
Type-2 plasticity loss	Fragmentation	Trending to ceiling	Unstable	Trend detection
Fine-tuning degradation	Either	Band shift	Band shift	Recalibration required
Quantization distortion	Either	Band shift	Potential decline	Recalibration check

Table 2: Summary of structural degradation pathologies, their energy and rank signatures, and the AI-7 detection mechanism. Each pathology produces a characteristic signature in the Dirichlet energy and/or effective rank diagnostics that AI-7 monitors continuously.

3 Geometric Foundations: The Dirichlet Energy Framework

The structural coherence invariant rests on a mathematical framework with three components: the Dirichlet energy functional from calculus on graphs, the spectral decomposition of the graph Laplacian that connects scalar energy to frequency-domain structure, and the adaptation of both to transformer architectures via the attention-as-dynamic-graph formulation. This section develops each component with the rigor required for a governance-grade specification.

3.1 The Dirichlet Energy Functional

3.1.1 Definition on Graphs

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ with $N = |\mathcal{V}|$ nodes, edge set $\mathcal{E}$, and edge weight function $W : \mathcal{E} \to \mathbb{R}_{>0}$. Let the adjacency matrix $A \in \mathbb{R}^{N \times N}$ encode edge weights: $A_{ij} = w_{ij}$ if $(i, j) \in \mathcal{E}$, and $A_{ij} = 0$ otherwise. Let $D = \mathrm{diag}(d_1, \ldots, d_N)$ be the degree matrix with $d_i = \sum_j A_{ij}$. The *combinatorial graph Laplacian* is $L = D - A$.

Given a node feature matrix $X \in \mathbb{R}^{N \times d}$, where each row $x_i \in \mathbb{R}^d$ is the feature vector of node i, the **Dirichlet energy** of the signal X with respect to $\mathcal{G}$ is:

$$\boxed{E_D(X) = \mathrm{Tr}(X^\top L X) = \sum_{(i,j)\in\mathcal{E}} w_{ij} \|x_i - x_j\|^2} \tag{1}$$

This equation admits a direct physical interpretation. The energy is a weighted sum of squared pairwise distances between connected nodes. It measures how much the signal X *varies* across the graph structure:

- If $E_D(X) = 0$ and the graph is connected, then $x_i = x_j$ for all pairs (i, j)—every node carries the identical signal. This is *global synchronization*, the state of zero information.
- If $E_D(X)$ is large, connected nodes carry very different signals—the signal is "rough" or "high-frequency" with respect to the graph topology.

The Dirichlet energy thus provides a scalar summary of representation health: too low indicates collapse (indistinguishability), too high indicates fragmentation (chaotic variation).

3.1.2 Normalized Variants

Two normalizations are standard in the literature, each with distinct properties:

Symmetric normalized Laplacian. Define $\hat{L} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$. The normalized Dirichlet energy is:

$$\hat{E}_D(X) = \mathrm{Tr}(X^\top \hat{L} X) = \sum_{(i,j)\in\mathcal{E}} \frac{w_{ij}}{\sqrt{d_i d_j}} \|x_i - x_j\|^2 \tag{2}$$

This normalization accounts for node degree, preventing high-degree nodes from dominating the energy. The eigenvalues of $\hat{L}$ lie in $[0, 2]$.

Random walk Laplacian. Define $L_{rw} = D^{-1}L = I - D^{-1}A$. The corresponding energy weights each edge by the transition probability from node i to node j, interpreting information propagation as a random walk on the graph.

AI-7 Energy Normalization Decision

Clause AI-7 specifies the **unnormalized (combinatorial) Dirichlet energy** (Equation 1) as the primary metric, with the symmetric normalized variant (Equation 2) as a secondary diagnostic. The rationale follows from Bison et al. (2024), who demonstrated that the normalized Dirichlet energy induced by the normalized Laplacian can reach zero even when representations are not fully identical—confusing "over-shrinking" (representations becoming small in norm) with "oversmoothing" (representations becoming indistinguishable). The unnormalized energy avoids this failure mode: it reaches zero if and only if all connected nodes have identical features.

The normalized variant remains valuable as a complementary diagnostic because it accounts for the degree distribution of the attention graph. When the attention pattern is highly non-uniform (some tokens receiving far more attention than others), the unnormalized energy is dominated by high-degree tokens. The normalized variant provides a degree-corrected view.

3.1.3 The Rayleigh Quotient Connection

For a single signal vector $x \in \mathbb{R}^N$ (one column of X), the *Rayleigh quotient* with respect to the Laplacian is:

$$R(x) = \frac{x^\top L x}{x^\top x} \tag{3}$$

The Rayleigh quotient satisfies $\lambda_1 \le R(x) \le \lambda_N$ for any nonzero x, where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of L. The minimum is achieved by the eigenvector corresponding to the smallest eigenvalue ($\lambda_1 = 0$ for the constant eigenvector on a connected graph), and the maximum by the eigenvector corresponding to λ_N.

This provides the spectral interpretation of the coherence band: a healthy representation maintains its Rayleigh quotient—and therefore its per-feature energy—in the *interior* of the spectral range, avoiding both the trivial constant solution (λ_1) and the maximally oscillatory solution (λ_N).

3.2 Spectral Graph Theory and the Laplacian Spectrum

The scalar energy $E_D(X)$ provides a summary, but the rigorous characterization of representational health requires the full spectral decomposition of the graph Laplacian.

3.2.1 The Eigendecomposition

Let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ be the eigenvalues of the combinatorial Laplacian L, with corresponding orthonormal eigenvectors $\phi_1, \phi_2, \ldots, \phi_N$. These eigenvectors form the *graph Fourier basis* (Bronstein et al., 2021): ϕ_1 is the constant vector (the "DC component"), and eigenvectors corresponding to larger eigenvalues represent progressively higher-frequency oscillations on the graph.

Any signal $x \in \mathbb{R}^N$ can be decomposed in this basis: $x = \sum_{k=1}^{N} \hat{x}_k \phi_k$, where $\hat{x}_k = \langle x, \phi_k \rangle$ are the graph Fourier coefficients. The Dirichlet energy decomposes as:

$$E_D(x) = x^\top L x = \sum_{k=1}^{N} \lambda_k \hat{x}_k^2 \tag{4}$$

This spectral decomposition is the foundation of the AI-7 frequency-domain diagnostics. It reveals that the Dirichlet energy is a *weighted sum* of squared Fourier coefficients, where each coefficient is weighted by its corresponding eigenvalue. High-frequency components (large λ_k) contribute disproportionately to the energy. Low-frequency components (small λ_k) contribute little.

3.2.2 Key Spectral Quantities

Three spectral quantities are central to the AI-7 diagnostic framework:

The Fiedler value (λ_2). The second-smallest eigenvalue of L, also known as the *algebraic connectivity*. A connected graph has $\lambda_2 > 0$; $\lambda_2 = 0$ if and only if the graph is disconnected. The Fiedler value measures the "tightest bottleneck" in the graph—the minimum energy required to partition the graph into two components. For representational health:

- A large λ_2 indicates strong global integration: information can propagate efficiently across the entire token sequence via the attention graph. The representation is coherent.
- A small λ_2 (approaching zero) indicates near-disconnection: the attention graph has fragmented into weakly connected subgraphs. Information cannot flow between parts of the sequence. The representation has lost global coherence.

The spectral gap (λ_2/λ_N). The ratio of the smallest nonzero eigenvalue to the largest eigenvalue. The spectral gap controls the rate at which message-passing operations (including self-attention) smooth the signal. A large spectral gap implies rapid convergence—which accelerates oversmoothing. A small spectral gap implies slow convergence but potential bottlenecks—which enables oversquashing. This is the *fundamental tension* identified by Topping et al. (2022): increasing the spectral gap to reduce oversquashing simultaneously accelerates oversmoothing.

The High-Frequency Energy Ratio (HFER). The fraction of total Dirichlet energy concentrated in the upper portion of the spectrum. Define a cutoff index $k^* = \lceil N/2 \rceil$ (or a task-specific threshold). Then:

$$\mathrm{HFER}(x) = \frac{\sum_{k=k^*}^{N} \lambda_k \hat{x}_k^2}{\sum_{k=1}^{N} \lambda_k \hat{x}_k^2} \tag{5}$$

The HFER captures the *distribution* of energy across the spectrum, not just its magnitude. A healthy representation has moderate total energy with most energy in low- and mid-frequency components (HFER < 0.5). An adversarially perturbed or fragmented representation shifts energy toward high frequencies (HFER > 0.5).

3.2.3 The Cheeger Inequality and Information Flow

The Cheeger inequality connects the spectral gap to the geometric bottleneck of the graph. Define the Cheeger constant (isoperimetric number) of $\mathcal{G}$ as:

$$h(\mathcal{G}) = \min_{S \subset \mathcal{V}, |S| \leq N/2} \frac{|\partial S|}{\min(\mathrm{vol}(S), \mathrm{vol}(\bar{S}))} \tag{6}$$

where ∂S is the set of edges crossing the partition and $\mathrm{vol}(S) = \sum_{i \in S} d_i$. Then the discrete Cheeger inequality states:

$$\frac{h^2}{2} \leq \lambda_2 \leq 2h \tag{7}$$

The governance implication is direct: a low Fiedler value ($\lambda_2 \approx 0$) implies a small Cheeger constant ($h \approx 0$), which means there exists a partition of the token set such that very little attention weight crosses the boundary. The attention mechanism has fragmented: one group of tokens is effectively processing in isolation from another. For reasoning tasks, this fragmentation can disconnect premises from conclusions; for dialogue tasks, it can disconnect context from response.

3.3 The Bridge to Transformers: Attention as a Dynamic Graph

3.3.1 The Attention Graph

In a transformer with H attention heads and L layers, each head h at layer ℓ computes an attention matrix $A^{(\ell,h)} \in \mathbb{R}^{T \times T}$, where T is the sequence length and:

$$A_{ij}^{(\ell,h)} = \mathrm{softmax}\left(\frac{q_i^{(\ell,h)\top} k_j^{(\ell,h)}}{\sqrt{d_k}} \right) \tag{8}$$

Each attention matrix defines a *weighted directed graph* over the token set: token i attends to token j with weight $A_{ij}^{(\ell,h)}$. Unlike GNNs, where the graph is fixed and given as input, the transformer's graph is *dynamic*—it is generated by the model itself, varies across layers and heads, and depends on the input.

For the purpose of Dirichlet energy computation, we require a symmetric structure. Define the *symmetrized attention graph* for head h at layer ℓ:

$$\tilde{A}^{(\ell,h)} = \frac{A^{(\ell,h)} + (A^{(\ell,h)})^\top}{2} \tag{9}$$

The corresponding Laplacian is $\tilde{L}^{(\ell,h)} = \tilde{D}^{(\ell,h)} - \tilde{A}^{(\ell,h)}$, where $\tilde{D}^{(\ell,h)}$ is the degree matrix of the symmetrized graph.

3.3.2 Attention-Weighted Dirichlet Energy

The **attention-weighted Dirichlet energy** at layer ℓ is defined as:

$$\boxed{E_{\mathrm{Attn}}^{(\ell)} = \frac{1}{H} \sum_{h=1}^{H} \mathrm{Tr}\left((X^{(\ell)})^\top \tilde{L}^{(\ell,h)} X^{(\ell)} \right) = \frac{1}{H} \sum_{h=1}^{H} \sum_{i,j} \tilde{A}_{ij}^{(\ell,h)} \|x_i^{(\ell)} - x_j^{(\ell)}\|^2} \tag{10}$$

where $X^{(\ell)} \in \mathbb{R}^{T \times d}$ is the hidden state matrix at layer ℓ and the average is taken over all H attention heads.

This formulation has several critical properties:

- **Architecture-aware:** The energy is computed with respect to the graph that the model *itself* generates. It measures the coherence of representations relative to the model's own attention structure, not an arbitrary external topology.
- **Head-averaged:** Averaging across heads provides robustness against single-head anomalies. Individual heads may specialize (some attending locally, some globally), but the average energy captures the aggregate geometric health.
- **Layer-indexed:** The energy is computed independently at each monitored layer, enabling layer-specific coherence bands that account for the natural energy evolution through the network depth.
- **Dynamic:** Because the attention graph changes with each input, the energy is input-dependent. This is a feature, not a limitation: it means AI-7 can detect input-specific representational anomalies (e.g., adversarial inputs that produce abnormal attention graphs).

3.3.3 The Energy of Attention: Physical Interpretation

The attention-weighted Dirichlet energy admits the following interpretation:

> $E_{\text{Attn}}^{(\ell)}$ measures the total "disagreement" between the model's attention structure and its representation structure at layer ℓ. If the model attends strongly between two tokens ($\tilde{A}_{ij}$ is large) but their representations are very different ($\|x_i - x_j\|$ is large), that pair contributes significant energy. A healthy model produces attention patterns that are *consistent* with its representations: tokens that attend to each other should have related—but not identical—representations.

When $E_{\text{Attn}}^{(\ell)} \to 0$: The attention mechanism is attending between tokens that have become identical. The attention is well-satisfied but trivially so—the model cannot distinguish between the tokens it connects. This is the *collapse* state.

When $E_{\text{Attn}}^{(\ell)}$ is very large: The attention mechanism connects tokens with wildly different representations. The model's attention structure and representation structure are misaligned—the model is "looking at" representations it cannot integrate coherently. This is the *fragmentation* state.

3.3.4 Alternative Graph Constructions

The symmetrized attention graph (Equation 9) is the primary graph for AI-7, but alternative constructions provide complementary information:

k-Nearest Neighbor (kNN) graph on hidden states. Construct a graph where each token is connected to its k nearest neighbors in the hidden state space (measured by cosine similarity or Euclidean distance). The Dirichlet energy on this graph measures representation smoothness with respect to *intrinsic similarity* rather than attention-defined connectivity. This is useful for detecting cases where the attention mechanism is healthy but the representations themselves have degenerated.

Residual stream graph. Construct a graph from the difference $X^{(\ell)} - X^{(\ell-1)}$, measuring the energy of the *update* at each layer rather than the representations themselves. A healthy network produces updates with moderate energy; zero-energy updates indicate the layer is not transforming the representations (the "curse of depth" in Pre-LN transformers), and high-energy updates indicate the layer is producing discontinuous changes.

AI-7 specifies the attention-weighted graph as mandatory and the kNN graph as an optional complementary diagnostic. The residual stream graph is recommended for extended monitoring in continuous learning deployments.

3.4 The DE Ratio: Layer-Wise Energy Dynamics

The Dirichlet energy at a single layer provides a snapshot. The *evolution* of energy across layers captures the dynamics of the network's geometric processing.

3.4.1 Definition

The **Dirichlet Energy Ratio** (DE Ratio) for layer ℓ is:

$$R_{DE}^{(\ell)} = \frac{E_{\text{Attn}}^{(\ell)}}{E_{\text{Attn}}^{(\ell-1)}} \tag{11}$$

where $E_{\text{Attn}}^{(0)}$ is computed on the initial embedding layer using a kNN graph (since no attention matrix exists at the input).

3.4.2 Interpretation

- $R_{DE}^{(\ell)} < 1$: Layer ℓ is a *smoother*—it reduces representation diversity. Self-attention layers are typically smoothers because they mix token information. A sustained sequence of $R_{DE} < 1$ across consecutive layers predicts oversmoothing.
- $R_{DE}^{(\ell)} \approx 1$: Layer ℓ *preserves* energy—it transforms representations without net smoothing or sharpening. This is the characteristic of well-balanced layers where attention mixing is counterbalanced by residual connections and feed-forward transformations.
- $R_{DE}^{(\ell)} > 1$: Layer ℓ is a *sharpener*—it increases representation diversity. Feed-forward layers and certain attention patterns can increase energy. Di Giovanni, Rowbottom, Chamberlain, Markovich, and Bronstein (2023) proved that graph convolutions with weight matrices having negative eigenvalues can induce sharpening, not just smoothing.

3.4.3 The Cumulative DE Product

The product of DE Ratios across depth gives the *cumulative energy scaling*:

$$\Pi_{DE}^{(L)} = \prod_{\ell=1}^{L} R_{DE}^{(\ell)} = \frac{E_{\text{Attn}}^{(L)}}{E_{\text{Attn}}^{(0)}} \tag{12}$$

A cumulative product approaching zero predicts imminent collapse. A cumulative product diverging predicts fragmentation. A stable product near 1.0 indicates energy-preserving dynamics across the full depth of the network—the ideal regime for the structural coherence invariant.

3.5 Effective Rank: The Complementary Diagnostic

3.5.1 Motivation

Zhang, Wei, Xu, and Liu (2025) proved that rank collapse is *strictly more general* than energy collapse: the effective rank of a representation matrix can decrease (representations lose dimensionality) without a corresponding decrease in Dirichlet energy. This occurs when representational diversity is lost in dimensions orthogonal to the graph structure—the surviving dimensions remain diverse along graph edges, preserving energy, but the overall representational capacity has diminished.

This theoretical result has a direct governance implication: *Dirichlet energy monitoring alone is insufficient to detect all forms of representational degradation.* A complementary metric is required.

3.5.2 Definition

The **effective rank** of a matrix $X \in \mathbb{R}^{T \times d}$ is defined via the Shannon entropy of the normalized singular values (Roy & Vetterli, 2007):

$$\operatorname{erank}(X) = \exp\left(-\sum_{i=1}^{\min(T,d)} \tilde{\sigma}_i \log \tilde{\sigma}_i\right) \tag{13}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots$ are the singular values of X and $\tilde{\sigma}_i = \sigma_i / \sum_j \sigma_j$ are the normalized singular values.

The effective rank has the following properties:

- $\operatorname{erank}(X) = 1$ if and only if X is rank-1 (all rows are scalar multiples of a single vector). This is the state of complete collapse.
- $\operatorname{erank}(X) = \min(T, d)$ if and only if all singular values are equal. This is the state of maximal representational diversity.
- The effective rank is scale-invariant and unitary-invariant—it depends only on the shape of the singular value distribution, not its magnitude or orientation.

3.5.3 The Effective Rank Floor

AI-7 specifies a minimum effective rank as a complementary diagnostic:

$$\operatorname{erank}(X^{(\ell)}) \geq r_{\min}(\ell) \tag{14}$$

where $r_{\min}(\ell)$ is calibrated from the validation distribution. The rank floor guards against dimensional collapse that energy monitoring misses. The combination of the Dirichlet energy band and the effective rank floor provides two-dimensional coverage of representational health:

State	Energy	Rank
Healthy	Within $[E^D_{\min}, E^D_{\max}]$	$\geq r_{\min}$
Oversmoothing	$< E^D_{\min}$	Low
Dimensional collapse	Within band (may appear normal)	$< r_{\min}$
Fragmentation	$> E^D_{\max}$	Variable
Adversarial perturbation	$> E^D_{\max}$	LID elevated

3.6 Formal Definition of Structural Coherence

With the preceding framework, the structural coherence property can be formally defined.

Definition: Structural Coherence

A model with L layers and monitored layer set $\mathcal{M} \subseteq \{1, \ldots, L\}$ is **structurally coherent** at time t if and only if:

1. **Energy band:** For every monitored layer $\ell \in \mathcal{M}$:

$$E^D_{\min}(\ell) \leq E^{(\ell)}_{\text{Attn}}(t) \leq E^D_{\max}(\ell) \tag{15}$$

2. **Rank floor:** For every monitored layer $\ell \in \mathcal{M}$:

$$\text{erank}(X^{(\ell)}(t)) \geq r_{\min}(\ell) \tag{16}$$

3. **Connectivity:** For every monitored layer $\ell \in \mathcal{M}$:

$$\lambda_2^{(\ell)}(t) \geq \tau_{\text{connected}} \tag{17}$$

where $\lambda_2^{(\ell)}$ is the Fiedler value of the symmetrized attention graph.

Violation of *any* of these three conditions at *any* monitored layer constitutes a structural coherence breach. The energy band is the primary invariant (carried in the MAI-1 `coherence-energy` field). The rank floor and connectivity threshold are complementary diagnostics (carried in extended fields).

This three-part definition ensures comprehensive coverage:

- The **energy band** catches oversmoothing, over-sharpening, and adversarial spectral disruption.
- The **rank floor** catches dimensional collapse invisible to energy monitoring.
- The **connectivity threshold** catches attention fragmentation—the Fiedler drop that indicates the model has lost global coherence in its processing.

Together, these three diagnostics certify that the model's internal geometry is neither collapsed (too smooth, too low-dimensional, or disconnected) nor fragmented (too rough, too noisy, or adversarially perturbed). The model is operating in the "Goldilocks zone" of representational health—diverse enough to represent complex inputs, smooth enough to generalize, and connected enough to reason globally.

4 The Lower Bound: Representational Collapse ($E^D_{\min}$)

The lower bound of the coherence band, $E^D_{\min}$, protects the model against representational collapse—the failure mode of excessive order, where the model simplifies its internal geometry to the point of information destruction. This section provides the formal characterization of collapse, its mathematical properties, the mechanisms that drive it, and the theoretical rates at which it occurs.

4.1 A Taxonomy of Collapse

The term "representational collapse" encompasses a family of related but distinct phenomena. A precise taxonomy is necessary because different collapse types produce different signatures in the AI-7 diagnostic framework and may require different intervention strategies.

Collapse Type	Definition	Energy Signature	Rank Signature
Complete collapse (Rank-0/1)	All representations converge to a single point or line	$E_D \to 0$	erank $\to 1$
Oversmoothing	Connected representations converge; disconnected may differ	$E_D \to 0$ (exponential)	erank declines progressively
Dimensional collapse	Representations span a low-dimensional subspace of the ambient space	May appear normal	erank $\ll d$
Neural collapse (healthy)	Last-layer features form Simplex ETF	Low at final layer	erank $\approx C$ (class count)
Neural collapse (pathological)	Intermediate layers form premature ETF	Low at intermediate layers	erank $\to C$ prematurely
Tunnel effect	Deep layers compress to task-minimum rank	Moderate	erank $\to C$ in deep layers; CKA $\to 1$ between layers
Anisotropic degeneration	Representations cluster in a narrow cone	May appear moderate	Effective rank moderate but cosine similarity high

Table 3: Taxonomy of representational collapse types. Each type is characterized by its energy and rank signatures. AI-7 uses the combination of energy band and rank floor to detect all types.

The key insight from this taxonomy is that Dirichlet energy alone detects complete collapse and oversmoothing but may miss dimensional collapse, the tunnel effect, and anisotropic degeneration. The effective rank floor (Equation 14) is necessary to close these gaps. The two diagnostics together provide comprehensive collapse detection.

4.2 Oversmoothing: The Canonical Collapse Pathway

4.2.1 The Graph Neural Network Theory

Oversmoothing is the most thoroughly characterized collapse mechanism. Li, Han, and Wu (2018) first identified it in graph convolutional networks: as depth increases, repeated application of the graph convolution operator—which is a form of Laplacian smoothing—forces all node representations to converge to a common value. The mechanism is intuitive: each layer averages a node's features with its neighbors' features, and iterated averaging converges to a global mean.

Oono and Suzuki (2020) provided the rigorous convergence bound. For a graph neural network with weight matrices $W^{(\ell)}$ and normalized adjacency $\hat{A}$, the representation at layer ℓ satisfies:

$$\|X^{(\ell)} - X^*\| \leq C_1 \exp(-C_2 \ell) \tag{18}$$

where X^* is the collapsed equilibrium (the projection onto the eigenspace of $\lambda_1 = 0$), C_1 depends on the initial representation, and C_2 depends on the spectral gap and the spectral norms of the weight matrices. Specifically, oversmoothing is guaranteed when:

$$\sigma_{\max}(W^{(\ell)}) \cdot (1 - \lambda_2(\hat{L})) < 1 \tag{19}$$

where $\sigma_{\max}(W^{(\ell)})$ is the largest singular value of the weight matrix and $\lambda_2(\hat{L})$ is the Fiedler value of the normalized Laplacian. This condition states that oversmoothing occurs when the weight matrices do not amplify features faster than the graph Laplacian smooths them.

4.2.2 The Dirichlet Energy Decay Rate

In the continuous diffusion limit, the Dirichlet energy decays at a rate governed by the first nonzero eigenvalue of the Laplacian:

$$E_D(X(t)) \leq E_D(X(0)) \cdot e^{-2\lambda_1 t} \tag{20}$$

where λ_1 is the first nonzero eigenvalue (i.e., λ_2 in the standard ordering) and t represents continuous "depth." For discrete GNN layers, Cai and Wang (2020) proved the analogous discrete bound: the energy after ℓ layers is bounded by $E_D(X^{(\ell)}) \leq \rho^\ell \cdot E_D(X^{(0)})$, where $\rho < 1$ depends on the product of the spectral norm of the weight matrices and the spectral radius of the normalized adjacency.

4.2.3 Empirical Decay Rates

The following quantitative results establish the severity of oversmoothing:

- Standard GCN on Cora: Dirichlet energy drops by *orders of magnitude* between layers 2 and 64. Performance degrades dramatically beyond approximately 4 layers.
- GCN, GAT, and GraphSAGE: exponential energy decay within 10–20 layers across all three architectures.
- Pure self-attention (Dong et al., 2021): rank collapse at *doubly exponential* rate $O(\exp(-\exp(L)))$—far more severe than GNN oversmoothing.

4.2.4 Mitigations and Their Limits

The literature has proposed numerous architectural interventions to combat oversmoothing. Their effectiveness is directly relevant to AI-7 because they define the landscape of energy behavior that the coherence band must accommodate:

- **PairNorm** maintains constant Dirichlet energy (≈ 1) by normalizing representations at each layer. However, performance still degrades after 32 layers despite constant energy—confirming Rusch et al.'s (2023) finding that constant energy is *necessary but not sufficient* for deep network performance.
- **GCNII** (Initial Residual + Identity Mapping) achieves both approximately constant energy and maintained performance. The initial residual connection preserves information from the input layer, preventing complete forgetting.
- **GraphCON** (Graph-Coupled Oscillator Networks) introduces a second-order ODE framework where zero-Dirichlet-energy steady states are *not exponentially stable*. This is a fundamental theoretical result: it means the collapse equilibrium is not an attractor in these dynamics, and the system can maintain nonzero energy indefinitely. On the Texas dataset, GraphCON-GCN achieves 85.4% versus GCN's 55.1%.

- **Residual connections in transformers** convert doubly-exponential rank collapse to power-law or logarithmic scaling, preventing catastrophic collapse but not eliminating the smoothing trend entirely.
- **DropEdge** randomly removes edges during training, slowing the decay rate but not changing the asymptotic behavior—the energy still converges to zero, just more slowly.

Honest Framing

The oversmoothing literature reveals a fundamental fact that AI-7 must acknowledge: *energy preservation alone does not guarantee healthy representations.* PairNorm maintains constant energy but performance degrades. This means the coherence band $[E^D_{\min}, E^D_{\max}]$ is a *necessary* condition for representational health, not a *sufficient* one. AI-7 certifies that the model has not entered a known-pathological geometric regime. It does not certify that the model's representations are optimal, well-calibrated, or semantically meaningful. This distinction between "not broken" and "provably good" is maintained throughout this specification.

4.3 Transformer Rank Collapse

4.3.1 The Doubly Exponential Result

Dong, Cordonnier, and Loukas (2021) proved the following result for pure self-attention:

Theorem (Dong et al.

Let $X^{(L)}$ be the output of an L-layer self-attention network without skip connections or feed-forward layers. Then:

$$\text{rank}(X^{(L)}) \to 1 \quad \text{at rate } O\left(\exp(-\exp(L))\right)$$

That is, the output converges to a rank-1 matrix—where all token representations are identical—at a *doubly exponential* rate in the number of layers.

This is the most severe convergence result in the oversmoothing literature. The doubly exponential rate means that even shallow pure-attention networks are close to rank-1: the collapse is not a concern limited to very deep architectures.

4.3.2 The Role of Architectural Components

The production transformers deployed in practice include several components that counteract this theoretical worst case:

Residual connections are the primary defense. By adding the input to the output at each layer ($X^{(\ell+1)} = X^{(\ell)} + \text{Attn}(X^{(\ell)})$), the residual stream maintains a "bypass" that preserves token identity. This converts the doubly exponential collapse to a much slower degradation. However, the residual stream itself can become dominant—creating the "curse of depth" where deeper layers contribute negligibly to the representation (the Pre-LN phenomenon).

Feed-forward networks (FFNs) act as nonlinear transformations that can increase representational diversity. In the energy framework of Di Giovanni et al. (2023), FFN layers can act as "sharpeners" ($R_{DE} > 1$), counterbalancing the smoothing effect of attention.

LayerNorm rescales representations but does *not* increase rank. As Dong et al. proved, right-multiplication by a matrix cannot increase rank—LayerNorm is a rank-preserving operation. Its contribution is to maintain the numerical stability of the signal, not to counteract collapse.

4.3.3 Empirical Evidence in Production Models

Shi et al. (2022) analyzed transformer oversmoothing from the graph perspective, finding that LayerNorm plays a critical role: if the standard deviation of the representations is sufficiently large, the transformer output converges to a specific low-rank subspace. Godey et al. (2024) demonstrated that anisotropy—where representations cluster in a narrow cone rather than occupying the full ambient space—is an *inherent* property of self-attention, observable across all transformer-based models regardless of training objective.

These findings establish that transformer rank collapse is not merely a theoretical concern about pure self-attention but a measurable phenomenon in deployed models, modulated (but not eliminated) by architectural components.

4.3.4 The Clause AI-7 Interpretation

AI-7 treats transformer rank collapse as a *runtime* risk, not merely an architectural design consideration. Even if a model's architecture includes residual connections, FFNs, and LayerNorm sufficient to prevent catastrophic collapse under normal conditions, the balance between smoothing and preserving forces can shift:

- Fine-tuning can alter the spectral norms of weight matrices, changing the smoothing rate.
- Input distributions that differ from training data can produce attention patterns with different spectral properties, altering the effective smoothing.
- Quantization can modify weight matrix spectra, shifting the smoothing/preserving balance.
- Continuous learning or online adaptation can progressively alter the dynamics.

The coherence band provides a continuous check on this balance. The energy floor $E^D_{\min}$ is calibrated from the model's energy profile on its validation distribution, representing the minimum energy observed under healthy conditions. A decline below this floor signals that the smoothing/preserving balance has shifted toward collapse.

4.4 Neural Collapse: Geometry of the Terminal Phase

4.4.1 The Four Properties

Papyan, Han, and Donoho (2020) documented four interconnected geometric properties that emerge during the terminal phase of training (after the training loss effectively reaches zero):

1. **NC1 (Variability Collapse):** Within-class variability vanishes: $\mathrm{Tr}(\Sigma_W)/\mathrm{Tr}(\Sigma_B) \to 0$, where Σ_W is the within-class covariance and Σ_B is the between-class covariance.
2. **NC2 (Convergence to Simplex ETF):** The class means converge to a Simplex Equiangular Tight Frame—a geometric structure that maximizes the pairwise angles between class centroids. For C classes, the cosine similarity between any two class means approaches $-1/(C-1)$.
3. **NC3 (Self-Duality):** The classifier weight vectors converge to the class means: $w_c \propto \mu_c$.
4. **NC4 (Nearest-Center Classification):** The classifier's decision rule becomes equivalent to assigning each input to its nearest class centroid.

Zhu et al. (2021) proved that this geometry is not merely an empirical observation but a mathematical necessity: cross-entropy loss with weight decay has a benign global landscape where the only global minimizers are Simplex ETF structures, and all other critical points are strict saddles.

4.4.2 Healthy vs. Pathological Collapse

The critical distinction for AI-7 is between neural collapse as a *healthy* terminal geometry and neural collapse as a *pathological* premature compression:

Healthy neural collapse occurs at the *last layer* of a classification network, *after* the earlier layers have developed rich, hierarchical representations. In this regime, NC is optimal: the last layer projects the high-dimensional feature space onto a maximally separable geometric structure. The Dirichlet energy at the last layer is low (features within each class are tightly clustered), but the energy at earlier layers remains high (diverse, expressive representations). Papyan et al. observed that convergence toward NC is associated with a median test accuracy improvement of 0.35–0.50% and a mean DeepFool adversarial robustness improvement of 0.2452.

Pathological neural collapse occurs in *intermediate layers*, or when the entire representation collapses to the ETF structure without first developing the hierarchical features needed for out-of-distribution generalization. This is the tunnel effect (Masarczyk et al., 2023): the deep layers compress the representation to rank $\approx C$ (number of classes), discarding the rich features learned by earlier layers. The tunneled representation passes task-specific benchmarks but fails on transfer, OOD detection, and complex reasoning.

4.4.3 Dirichlet Energy Characterization of Neural Collapse

In the energy framework, neural collapse produces a distinctive layer-wise signature:

- **Healthy NC:** Energy is high in early and middle layers (rich representation), declining only in the final layer(s) where the ETF projection occurs. The effective rank in middle layers remains high ($\gg C$).
- **Pathological NC:** Energy drops *early*—at intermediate layers well before the classifier. The effective rank in intermediate layers approaches C. The DE Ratio is consistently < 1 across a long sequence of layers, indicating progressive compression throughout the depth.
- **Linguistic collapse (Wu & Papyan, 2024):** In LLMs, where the vocabulary size far exceeds the embedding dimension and the task structure is autoregressive rather than classification, partial NC manifests as an anisotropic concentration of embedding space. Long-tail tokens lose representational resolution while high-frequency tokens maintain separation.

AI-7's layer-specific coherence band naturally distinguishes these cases: $E^D_{\min}(\ell)$ is calibrated independently for each monitored layer, so a low-energy final layer (healthy NC) does not trigger a floor violation, while a low-energy intermediate layer (pathological NC) does.

4.5 Dimensional Collapse: The Energy-Invisible Degradation

4.5.1 The Formal Characterization

Dimensional collapse is the most subtle form of representational degradation because it can be *invisible* to Dirichlet energy monitoring. Zhang et al. (2025) proved the key theoretical result:

Theorem (Zhang et al.

Rank collapse is strictly more general than energy collapse. There exist representation matrices X with:

- Low effective rank ($\text{erank}(X) \ll d$)
- Moderate Dirichlet energy ($E^D_{\min} \leq E_D(X) \leq E^D_{\max}$)

This occurs when the representation diversity is lost in dimensions *orthogonal* to the graph structure: the surviving dimensions remain diverse along graph edges (maintaining energy) but the overall representational capacity has diminished (low rank).

This result is the formal justification for AI-7's dual-diagnostic architecture. Dirichlet energy measures spatial smoothness *along the graph*; effective rank measures dimensionality *in the ambient space*. A model can lose ambient dimensionality while preserving smoothness variation—the energy looks healthy but the representation is impoverished.

4.5.2 Mechanisms of Dimensional Collapse

Jing et al. (2022) identified two root causes in self-supervised learning that generalize to the broader foundation model regime:

Augmentation-induced rank reduction. Strong data augmentation during training makes the covariance of the augmented views low-rank. The model learns representations that are invariant to the augmentation—which is the goal—but if the augmentations are too aggressive, the invariant subspace is much smaller than the ambient dimension, and the representation collapses onto it.

Implicit low-rank regularization. Over-parameterized networks trained with weight decay exhibit an implicit bias toward low-rank solutions. Feng, Zheng, Huang, Zhao, Jordan, and Zha (2022) proved a universal monotone decreasing property of network rank: under differential and algebraic composition rules, deeper networks produce progressively lower-rank representations. On ImageNet, ResNet-50 (with 2048-dimensional final representations) requires only ∼5–10% of principal components for 95% classification accuracy—indicating that the effective dimensionality is far below the ambient dimension.

Batch normalization as a partial remedy. Daneshmand, Kohler, Bach, Hofmann, and Lucchi (2020) proved that vanilla deep networks converge to rank-1 in depth, but batch normalization preserves rank $\Omega(\sqrt{d})$ for width d. For a network with width 32, vanilla networks drop to rank ∼1 by depth 20–30, while batch-normalized networks maintain rank $\geq$∼5.7. This result establishes that normalization layers provide some rank preservation but not full-rank maintenance.

4.5.3 The Effective Rank Profile

Ansuini, Laio, Macke, and Zoccolan (2019) provided the first comprehensive study of intrinsic dimensionality across CNN layers using the Two-NN estimator. They found a characteristic "hunchback" profile: intrinsic dimension increases in early layers, peaks in the middle, and progressively decreases in final layers. For VGG-16, intrinsic dimension peaked at ∼30–40 in early layers and dropped to ∼5–10 in final layers (versus an embedding dimension of 4096). Crucially, networks trained on random labels showed *expanding* intrinsic dimension—no compression—confirming that the hunchback profile is a signature of meaningful learning, not a computational artifact.

Valeriani et al. (2023) extended this analysis to large transformers (ESM-2, iGPT), finding a related pattern: early peak → long plateau → final ascent. For ESM-2 (3B parameters), intrinsic dimension peaked at ∼60–80 in early layers and dropped to a plateau of ∼15–20 in middle layers.

4.5.4 The Clause AI-7 Interpretation

The effective rank floor $r_{\min}(\ell)$ is calibrated from these empirical profiles. For each monitored layer, $r_{\min}(\ell)$ is set to a fraction of the effective rank observed on the validation distribution—typically the 1st percentile to provide a conservative floor. A drop below $r_{\min}(\ell)$ at any monitored layer signals dimensional collapse, even if the Dirichlet energy remains within the certified band.

The layer-specific nature of the rank floor is essential: the hunchback profile means that different layers naturally operate at different effective ranks. A rank of 15 at a middle layer might be healthy for a 3B-parameter model, while a rank of 15 at an early layer would indicate severe compression. The calibration procedure (Section 8) accounts for this variation.

4.6 The Oversmoothing–Oversquashing Tension

Topping et al. (2022) and Giraldo et al. (2022) identified a fundamental tension that constrains the collapse floor:

The Spectral Gap Tradeoff

Increasing the spectral gap (the ratio λ_2/λ_N) of the graph Laplacian:

- **Reduces oversquashing:** A larger λ_2 (stronger connectivity) means information can flow more easily across the graph, reducing bottleneck compression.
- **Accelerates oversmoothing:** A larger spectral gap increases the rate of Dirichlet energy decay (Equation 20), pushing representations toward collapse faster.

These two pathologies cannot be alleviated simultaneously through the spectral gap alone. Structural modifications (graph rewiring, architectural changes) are required to address both.

For AI-7, this tension means the coherence band cannot be arbitrarily narrow. The model must operate in a regime where the spectral gap is large enough to avoid oversquashing (Section 5) but not so large that oversmoothing is accelerated beyond the architecture's ability to counteract it. The width of the coherence band $[E^D_{\min}, E^D_{\max}]$ reflects this fundamental tension: a wider band accommodates the natural variation in energy that arises from the tradeoff, while the floor and ceiling guard against the pathological extremes.

4.7 Summary: The Collapse Floor

The collapse floor $E^D_{\min}$ guards against a family of related degradation pathologies:

1. **Complete collapse** ($E_D \to 0$, erank $\to 1$): Detected by energy floor violation.
2. **Oversmoothing** (E_D decays exponentially, erank declines): Detected by sustained $R_{DE} < 1$ and energy approaching the floor.
3. **Dimensional collapse** (E_D may appear normal, erank $\ll d$): Detected by rank floor violation.
4. **Pathological neural collapse** (E_D drops at intermediate layers): Detected by layer-specific energy floor violation.
5. **Tunnel effect** (erank $\to C$ in deep layers): Detected by layer-specific rank floor violation.
6. **Type-1 plasticity loss** (E_D trends toward floor over time): Detected by temporal trend analysis on the energy trace.

The common thread across all collapse types is information destruction: the model loses representational capacity that cannot be recovered without retraining or architectural intervention. The collapse floor provides early warning—detecting the geometric precursors of failure before they propagate to output-level degradation.

5 The Upper Bound: Representational Fragmentation ($E^D_{\max}$)

The upper bound of the coherence band, $E^D_{\max}$, protects the model against representational fragmentation—the failure mode of excessive disorder, where the model's internal geometry becomes chaotically sensitive, topologically bottlenecked, or spectrally noisy. If collapse is the pathology of "too much order," fragmentation is the pathology of "too much chaos." This section formally characterizes the fragmentation failure mode, the mechanisms that drive it, and the spectral signatures that AI-7 monitors at the ceiling.

5.1 Oversquashing: Information Compression at Topological Bottlenecks

5.1.1 The Phenomenon

Alon and Yahav (2021) identified oversquashing as a fundamental limitation of message-passing architectures: when a model must propagate information from a source node to a distant target node through a bottleneck in the graph topology, the information from an exponentially growing receptive field is compressed into a fixed-size vector. The gradient signal from the target to the source decays exponentially with distance, making long-range dependencies unlearnable.

Formally, the sensitivity of node v's representation at layer m to node u's input features is bounded by:

$$\left\| \frac{\partial h_v^{(m)}}{\partial x_u} \right\| \leq C \cdot (\hat{A}^m)_{vu} \tag{21}$$

where $\hat{A}$ is the normalized adjacency and $(\hat{A}^m)_{vu}$ is the (v, u)-entry of $\hat{A}$ raised to the m-th power. When the graph has bottlenecks between v and u, this entry decays exponentially, and information from u cannot influence v's representation regardless of how many layers are applied.

5.1.2 The Energy Interpretation

Oversquashing is a *high-energy* pathology, though this may seem counterintuitive. The mechanism operates as follows:

- The graph has bottleneck edges—narrow passages through which information must flow between regions of the graph.
- Nodes on opposite sides of the bottleneck develop *different* representations because information exchange between the regions is throttled. Each region evolves semi-independently.
- The attention weights across the bottleneck may be large (the model "tries" to connect the regions), but the representations are very different (the connection fails to integrate them).
- The Dirichlet energy contribution from bottleneck edges is therefore large: high attention weight $\tilde{A}_{ij}$ multiplied by large representation difference $\|x_i - x_j\|^2$.

The result is a paradox: the model is in a high-energy state (large representation differences across attended edges) but is *informationally poor* (it cannot capture long-range dependencies). The energy is concentrated at bottleneck edges rather than distributed across the graph, creating a characteristic spectral signature.

5.1.3 Topology as the Primary Determinant

Di Giovanni, Giusti, Barbero, Luise, Liò, and Bronstein (2023) established three main results about the relative importance of width, depth, and topology in determining oversquashing:

1. **Width mitigates but does not resolve:** Increasing hidden dimension reduces oversquashing but simultaneously increases sensitivity to *all* inputs, not just the relevant ones. The model becomes uniformly more responsive rather than selectively more connected.

2. **Depth cannot help:** Adding layers does not resolve oversquashing because the gradient vanishing through bottlenecks dominates. The oversquashing bound (Equation 21) decays with depth regardless of layer count.

3. **Topology plays the greatest role:** The *commute time* (or equivalently, the *effective resistance*) between nodes determines oversquashing severity. Black, Wan, Nayyeri, and Wang (2023) formalized this via the total effective resistance: $\text{Res}_{\mathcal{G}} = \sum_{v,u} R(v,u)$, which provides a global oversquashing measure with tighter bounds than the spectral gap alone.

5.2 Curvature Analysis: The Geometric Origin of Bottlenecks

5.2.1 Ricci Curvature on Graphs

Topping, Di Giovanni, Chamberlain, Dong, and Bronstein (2022) provided the geometric explanation for oversquashing by connecting it to the *curvature* of the graph.

In Riemannian geometry, curvature measures how geodesics converge or diverge. On graphs, the discrete analogue—Ollivier-Ricci curvature—measures whether the neighborhoods of two connected nodes are "close together" (positive curvature) or "far apart" (negative curvature):

- **Positive curvature:** The neighborhoods of nodes i and j overlap significantly. Information flows easily between them. These edges are "bridges" that facilitate integration.

- **Negative curvature:** The neighborhoods of nodes i and j are disjoint or nearly so. Information must squeeze through the edge (i,j) to pass between the two neighborhoods. These are *bottleneck edges* that cause oversquashing.

Topping et al. proved that negatively curved edges, measured via **Balanced Forman Curvature** (BFC), are the geometric cause of oversquashing. BFC provides a computationally tractable lower bound on Ollivier-Ricci curvature:

$$\text{Ric}_F(i,j) = \frac{2}{d_i} + \frac{2}{d_j} - 2 + \frac{2|\#\triangle(i,j)|}{\max(d_i, d_j)} + \text{corrections} \tag{22}$$

where d_i, d_j are node degrees and $\#\triangle(i,j)$ counts shared triangles. Edges with $\text{Ric}_F(i,j) < 0$ are bottlenecks. The computational complexity of BFC is $O(|\mathcal{E}| \cdot D^2)$, where D is the maximum degree—significantly cheaper than the $O(|\mathcal{E}| \cdot D^3)$ required for exact Ollivier-Ricci curvature.

5.2.2 The Unified Curvature Framework

Nguyen et al. (2023) provided a unified framework connecting both failure modes through curvature:

Curvature–Pathology Correspondence

- **Positive Ollivier-Ricci curvature $\longleftrightarrow$ Oversmoothing.** Positively curved regions of the graph facilitate information exchange so effectively that representations converge. Energy flows *out* of the signal and into the graph structure.
- **Negative Ollivier-Ricci curvature $\longleftrightarrow$ Oversquashing.** Negatively curved edges create bottlenecks that compress information, forcing representations on opposite sides to diverge. Energy accumulates at bottleneck edges.

Healthy representations require a *balanced* curvature distribution: enough positive curvature for integration, enough structural diversity to avoid global smoothing.

This curvature framework provides a geometric interpretation of the AI-7 coherence band: the band bounds correspond to the range of curvature distributions consistent with healthy information flow. Too much positive curvature drives the system below the floor (oversmoothing); too much negative curvature drives it above the ceiling (oversquashing/fragmentation).

5.3 Over-Sharpening: The High-Energy Mirror of Oversmoothing

5.3.1 The Theoretical Result

Di Giovanni, Rowbottom, Chamberlain, Markovich, and Bronstein (2023) proved a result that fundamentally extends the energy framework beyond the traditional oversmoothing narrative: graph convolutions do *not* always decrease Dirichlet energy. When the weight matrices have *negative eigenvalues*, the convolution operator acts as a *high-pass filter*, amplifying high-frequency components and increasing energy.

Specifically, the energy change at layer ℓ depends on the spectral properties of the weight matrix $W^{(\ell)}$:

- If $W^{(\ell)}$ has only positive eigenvalues: the layer smooths (low-pass filter), $R_{DE}^{(\ell)} < 1$.
- If $W^{(\ell)}$ has negative eigenvalues: the layer sharpens (high-pass filter), $R_{DE}^{(\ell)} > 1$.
- The net effect depends on the alignment between the weight matrix spectrum and the graph Laplacian spectrum.

5.3.2 Over-Sharpening as a Failure Mode

Over-sharpening occurs when the cumulative effect of high-pass filtering across layers amplifies high-frequency components to the point where the representation becomes dominated by noise:

$$\Pi_{DE}^{(L)} = \prod_{\ell=1}^{L} R_{DE}^{(\ell)} \gg 1 \tag{23}$$

In this regime, the model's representations are "rough"—neighboring tokens in the attention graph have very different feature vectors, not because they encode different semantic content, but because high-frequency spectral noise has been progressively amplified through depth.

5.3.3 The Clause AI-7 Interpretation

Over-sharpening establishes the theoretical basis for the energy ceiling $E_{\max}^{D}$. While oversmoothing (the floor pathology) results from excessive low-pass filtering, over-sharpening (the ceiling pathology) results from excessive high-pass filtering. The two pathologies are *dual*: they occupy opposite ends of the Dirichlet energy spectrum and produce opposite spectral signatures. The coherence band $[E_{\min}^{D}, E_{\max}^{D}]$ guards against both.

The DE Ratio diagnostic is particularly informative for detecting over-sharpening: a sustained sequence of $R_{DE}^{(\ell)} > 1$ across multiple layers, or a single layer with $R_{DE}^{(\ell)} \gg 1$, signals that the model is in a sharpening regime. The cumulative DE product (Equation 12) diverging above 1.0 predicts ceiling violation.

5.4 Chaotic Dynamics and Type-2 Plasticity Loss

5.4.1 The Dynamical Systems Perspective

Neural networks can be viewed as discrete dynamical systems: each layer maps the representation from one state to the next. The stability of these dynamics is characterized by *Lyapunov exponents*—the rate at which infinitesimally close trajectories converge or diverge.

In the context of plasticity loss, Dohare et al. (2024) and the associated literature distinguish two dynamical regimes:

- **Type-1 (Contractive):** Negative Lyapunov exponents. Nearby trajectories converge. The system is stable but rigid—incapable of differentiating between similar inputs because they all collapse to the same representation. This is the collapse regime (Section 4).
- **Type-2 (Expansive/Chaotic):** Positive Lyapunov exponents. Nearby trajectories diverge exponentially. The system is unstable—similar inputs map to vastly different representations, destroying the smooth interpolation that enables generalization.

5.4.2 The Energy Signature of Chaos

In the Dirichlet energy framework, chaotic dynamics produce a distinctive signature:

- **Energy explosion:** Because similar inputs diverge in the representation space, the pairwise distances between tokens that should be similar (and are connected by attention) become large. The Dirichlet energy increases rapidly with depth.
- **High variance:** The energy is not just high but *unstable*—it fluctuates dramatically across inputs and across layers. The variance of $E_{\text{Attn}}^{(\ell)}$ across a batch of inputs is a secondary diagnostic for chaotic dynamics.
- **HFER elevation:** Chaotic dynamics amplify high-frequency components preferentially, because high-frequency perturbations are exactly the ones that diverge fastest. The HFER increases as the chaos develops.

5.4.3 The Clause AI-7 Interpretation

Type-2 plasticity loss is a *progressive* pathology: the model transitions from stable to chaotic dynamics over time, typically during continuous learning or extended fine-tuning. AI-7's continuous monitoring detects this transition as it develops. The energy ceiling $E_{\max}^{D}$ catches the terminal state (energy above the certified maximum), while temporal trend analysis on the energy trace detects the progression—a monotonically increasing energy trend across batches signals an approaching ceiling violation.

5.5 Adversarial Vulnerability and the Fragmentation Ceiling

5.5.1 The Representation Geometry of Adversarial Robustness

Five independent research threads converge on the conclusion that representational geometry is a primary determinant of adversarial robustness:

Smoother manifolds produce more robust models. Engstrom, Ilyas, Santurkar, Tsipras, Tran, and Madry (2019) demonstrated that adversarially trained models produce perceptually aligned, approximately invertible representations—their internal manifolds are smoother than those of standard models. Salman et al. (2020) showed that robust ImageNet models transfer better to 12 downstream datasets, indicating that the smooth geometry induced by adversarial training produces more generally useful representations.

Spectral uniformity correlates with robustness. Cheng, Zhu, Zhang, and Liu (2022) demonstrated through Feature Spectral Regularization (FSR) that eigenvectors corresponding to *smaller* eigenvalues of the representation covariance matrix are more non-robust—adversarial perturbations preferentially add components along these low-eigenvalue directions. Penalizing the dominance of the largest eigenvalue (spreading the eigenvalue distribution more uniformly) improves robustness.

Lower effective dimensionality correlates with robustness. Khachaturov et al. (2024) found a near-linear inverse relationship between effective dimensionality and adversarial robustness across ResNet, ShuffleNet, and YOLO architectures on CIFAR-10/100 and ImageNet. Adversarial training methods (PGD-AT, TRADES, AWP) all reduce effective dimensionality—the model learns to concentrate its representations on a lower-dimensional manifold that is harder to perturb off of.

Adversarial regions have elevated local intrinsic dimensionality. Ma et al. (2018) demonstrated that the regions of the representation space occupied by adversarial examples have significantly higher local intrinsic dimensionality (LID) than regions occupied by clean data, with detection AUC exceeding 0.90 across five different attack strategies (FGSM, BIM, JSMA, DeepFool, C&W). The adversarial perturbation pushes the representation into a high-dimensional neighborhood—a "rough" region of the manifold where interpolation is unreliable.

Lipschitz constant bounds energy change. Cissé, Bojanowski, Grave, Dauphin, and Usunier (2017) showed through Parseval Networks that constraining the Lipschitz constant of each layer to ≤ 1 (by maintaining weight matrices as approximate Parseval tight frames) improves adversarial robustness. The Lipschitz constant directly bounds how much the Dirichlet energy can change under input perturbation: if the Lipschitz constant of the network is κ, then an input perturbation of magnitude ϵ can change the representation distances by at most $\kappa\epsilon$, bounding the energy change.

5.5.2 The Adversarial–Energy Connection

These five threads converge on a unified picture:

The Adversarial–Energy Hypothesis

A model whose internal representations maintain Dirichlet energy within the certified coherence band $[E^D_{\min}, E^D_{\max}]$ is more robust to adversarial perturbation than one whose energy is unbounded, because:

1. The energy floor prevents *collapsed* representations, which are vulnerable because the collapse direction provides a low-dimensional subspace that adversaries can exploit.
2. The energy ceiling prevents *fragmented* representations, which are vulnerable because chaotic sensitivity amplifies perturbations.
3. The effective rank floor ensures that representations are not concentrated on a subspace that adversaries can target with aligned perturbations.
4. The connectivity threshold ensures that the attention graph maintains global integration, preventing adversarial inputs from isolating parts of the representation.

Honest Framing

The adversarial–energy hypothesis is supported by strong convergent evidence from multiple independent research groups but has not been formally proven as a certifiable guarantee. AI-7 does not claim that maintaining the coherence band provides a *guarantee* of adversarial robustness. It claims that coherence band violation is a *necessary condition* for certain classes of representational vulnerability: a model outside the band is certifiably in a geometric regime associated with adversarial fragility. A model within the band has not been proven safe from all adversarial attacks, but its representational geometry is in the regime empirically associated with greater robustness. This is the distinction between a detection signal and a safety guarantee.

5.6 The Fiedler Drop: Attention Graph Disconnection

5.6.1 The Diagnostic

Section 3.2.2 introduced the Fiedler value (λ_2) as a measure of algebraic connectivity. In the context of the fragmentation ceiling, the Fiedler value provides a specific and powerful diagnostic: attention graph disconnection.

Under adversarial stress, hallucinatory states, or distributional shift, the attention graph often fractures into isolated subgraphs. The model's attention mechanism stops integrating global context and instead attends only within local neighborhoods. In the spectral domain, this manifests as a sharp drop in λ_2—the graph approaches disconnection.

5.6.2 Combined Diagnostic

The Fiedler drop is most diagnostic when combined with the energy ceiling:

Energy	Fiedler λ_2	HFER	Interpretation
High ($> E^D_{\max}$)	Low (dropping)	High	Adversarial attack or hallucinatory fragmentation
High ($> E^D_{\max}$)	Normal	High	Over-sharpening or chaotic dynamics
Normal	Low (dropping)	Normal	Attention fragmentation without energy spike
Normal	Normal	Normal	Healthy
Low ($< E^D_{\min}$)	Low (dropping)	Low	Severe collapse with disconnection

The combination of high energy, low Fiedler value, and elevated HFER is the most dangerous signature: it indicates that the model's attention has fragmented *and* the representations in the disconnected subgraphs are diverging. This is the spectral signature of a model that has lost coherent reasoning—different parts of the input are being processed independently, with no integration of global context, and the isolated processes are producing incompatible representations.

5.7 The Attention Entropy Connection

5.7.1 Entropy as a Complementary View

Zhai et al. (2023) established that attention entropy provides a complementary perspective on the oversmoothing–oversquashing spectrum:

$$H(\text{Attn}) \geq \log(T) - \sigma \cdot \sqrt{T} + O(1) \tag{24}$$

where T is the sequence length and σ is the spectral norm of the attention logits. The bound reveals:

- **High attention entropy** ($H \to \log T$): Uniform attention. Every token attends equally to every other token. This is the maximum-mixing regime that drives oversmoothing—a low-energy pathology.
- **Low attention entropy** ($H \to 0$): Concentrated attention. Each token attends to a single other token. This creates sparse, potentially bottlenecked attention graphs—a regime conducive to oversquashing and fragmentation.
- **Moderate attention entropy**: Selective but distributed attention. Tokens attend preferentially but maintain diverse connections. This is the regime consistent with healthy Dirichlet energy.

5.7.2 The Clause AI-7 Interpretation

Attention entropy is not a primary AI-7 diagnostic (Dirichlet energy subsumes the information it provides in most cases), but it serves as an efficient proxy for preliminary screening. The computational cost of attention entropy is $O(T \cdot H)$ per layer—negligible compared to the Dirichlet energy computation. When attention entropy falls outside a healthy range, it triggers full spectral analysis on that layer.

The attention entropy bound (Equation 24) also provides a connection to the spectral norm of the attention logits, which is itself related to the Lipschitz constant of the attention mechanism. Bounding the spectral norm bounds the attention entropy, which bounds the rate of Dirichlet energy change—closing the theoretical loop between adversarial robustness (Lipschitz bounds), attention dynamics (entropy), and representational geometry (Dirichlet energy).

5.8 Summary: The Fragmentation Ceiling

The fragmentation ceiling $E^D_{\max}$ guards against a family of high-energy pathologies:

1. **Oversquashing** (high energy at bottleneck edges, information compression): Detected by energy ceiling violation combined with Fiedler value analysis and effective resistance computation.
2. **Over-sharpening** (high-pass filtering amplifying high-frequency components): Detected by sustained $R_{DE} > 1$ and cumulative DE product divergence.
3. **Type-2 plasticity loss / chaotic dynamics** (positive Lyapunov exponents, exponential divergence): Detected by energy ceiling violation with high variance and temporal trend analysis.
4. **Adversarial perturbation** (high-frequency noise injection): Detected by energy spike, HFER elevation, and LID increase.
5. **Attention graph disconnection** (Fiedler drop): Detected by λ_2 falling below the connectivity threshold $\tau_{\text{connected}}$.

The common thread across all fragmentation pathologies is *information incoherence*: the model's internal representations are no longer organized in a way that supports reliable computation. The representations may be high-dimensional (high rank) and energetic (high Dirichlet energy), but the energy is noise rather than signal—it encodes chaotic variation rather than meaningful semantic structure. The fragmentation ceiling provides the upper bound that separates "energetic and expressive" from "energetic and broken."

6 Mathematical Formalization of Clause AI-7

This section translates the geometric foundations (Section 3) and the failure mode characterizations (Sections 4–5) into the formal invariant specification that populates the MAI-1 `coherence-energy` field. Every parameter, threshold, and transition rule is defined with the precision required for binary compliance determination.

6.1 The Structural Coherence Invariant

6.1.1 Primary Invariant Statement

The Clause AI-7 invariant is:

Mandatory Invariant: Structural Coherence (AI-7)

At every attested measurement point, for every monitored layer $\ell \in \mathcal{M}$:

$$\boxed{E^D_{\min}(\ell) \leq E^{(\ell)}_{\text{Attn}} \leq E^D_{\max}(\ell)} \tag{25}$$

where:

- $E^{(\ell)}_{\text{Attn}}$ is the attention-weighted Dirichlet energy at layer ℓ (Equation 10).
- $E^D_{\min}(\ell)$ is the certified collapse floor for layer ℓ.
- $E^D_{\max}(\ell)$ is the certified fragmentation ceiling for layer ℓ.
- $\mathcal{M} \subseteq \{1, \ldots, L\}$ is the monitored layer set.

This invariant **SHALL** be evaluated at the monitoring frequency specified by the deployment context (Section 6.5). Violation of the invariant at *any* monitored layer constitutes a structural coherence breach.

6.1.2 Complementary Invariant: Effective Rank Floor

Complementary Invariant: Effective Rank (AI-7-R)

At every attested measurement point, for every monitored layer $\ell \in \mathcal{M}$:

$$\text{erank}(X^{(\ell)}) \geq r_{\min}(\ell) \tag{26}$$

where:

- $\text{erank}(X^{(\ell)})$ is the effective rank of the hidden state matrix at layer ℓ (Equation 13).
- $r_{\min}(\ell)$ is the certified rank floor for layer ℓ.

The effective rank invariant is **complementary**: it **SHALL** be evaluated alongside the primary energy invariant but is carried in an extended MAI-1 field (`coherence-rank`) rather than the primary `coherence-energy` field. Violation of the rank floor constitutes a structural coherence breach even if the energy band is satisfied.

6.1.3 Complementary Invariant: Attention Connectivity

Complementary Invariant: Fiedler Connectivity (AI-7-λ)

At every attested measurement point, for every monitored layer $\ell \in \mathcal{M}$:

$$\bar{\lambda}_2^{(\ell)} \geq \tau_{\text{connected}} \tag{27}$$

where:

- $\bar{\lambda}_2^{(\ell)} = \frac{1}{H}\sum_{h=1}^{H} \lambda_2(\tilde{L}^{(\ell,h)})$ is the head-averaged Fiedler value of the symmetrized attention graph at layer ℓ.
- $\tau_{\text{connected}}$ is the certified connectivity threshold.

The Fiedler invariant is **complementary**: it detects attention graph fragmentation that may accompany or precede energy ceiling violations. Violation indicates that the model's attention mechanism has lost global integration capacity.

6.1.4 The Composite Coherence Predicate

The three invariants combine into a single composite predicate:

$$\mathcal{SC}(\ell, t) = \mathbb{I}\Big[E^D_{\min}(\ell) \leq E^{(\ell)}_{\text{Attn}}(t) \leq E^D_{\max}(\ell)\Big] \wedge \mathbb{I}\Big[\text{erank}(X^{(\ell)}(t)) \geq r_{\min}(\ell)\Big] \wedge \mathbb{I}\Big[\bar{\lambda}_2^{(\ell)}(t) \geq \tau_{\text{connected}}\Big] \tag{28}$$

The model is structurally coherent at time t if and only if:

$$\forall \ell \in \mathcal{M}: \quad \mathcal{SC}(\ell, t) = 1 \tag{29}$$

6.2 The Monitored Layer Set

6.2.1 Layer Selection Rationale

Not every layer requires monitoring. The phenomenology of structural degradation (Section 2) and the empirical literature on intrinsic dimensionality profiles (Section 4.5) establish that different layers serve different roles and exhibit different failure signatures:

- **Early layers** ($\ell \approx L/4$): Where the initial representation is formed. Collapse here indicates that the model has failed to develop rich features from the input—the "extractor" phase (Masarczyk et al., 2023) is compromised.
- **Middle layers** ($\ell \approx L/2$): Where collapse typically *initiates*. The transition from extractor to tunnel occurs in middle layers, and oversmoothing accumulates here. Middle layers are the highest-priority monitoring targets.
- **Deep layers** ($\ell \approx 3L/4$): Where the tunnel effect manifests (rank drops to $\approx C$). These layers are critical for detecting fine-tuning-induced compression.
- **Penultimate layer** ($\ell = L - 1$): Where fragmentation impacts output. High energy or Fiedler drops here directly precede output-level failure.

6.2.2 The Mandatory Monitoring Set

Monitored Layer Set

Clause AI-7 specifies the following **mandatory** monitored layer set:

$$\mathcal{M} = \left\{ \left\lfloor \frac{L}{4} \right\rfloor, \left\lfloor \frac{L}{2} \right\rfloor, \left\lfloor \frac{3L}{4} \right\rfloor, L-1 \right\} \tag{30}$$

where L is the total number of transformer layers.
Implementations **MAY** monitor additional layers beyond the mandatory set. Implementations **MUST NOT** monitor fewer than the four mandatory layers. For models with $L < 8$, the monitored set **SHALL** include at minimum the first, middle, and last layers.

This four-layer monitoring set is consistent with the empirical finding from Feng et al. (2022) that monitoring the first, 25%, 50%, 75%, and last layers captures the monotonic rank decline pattern. The four-layer set balances diagnostic coverage against computational overhead.

6.3 The Dirichlet Energy Computation Specification

6.3.1 Input Specification

At each monitored layer $\ell \in \mathcal{M}$, the following quantities are required:

- $X^{(\ell)} \in \mathbb{R}^{T \times d}$: The hidden state matrix. Row i is the d-dimensional representation of token i after layer ℓ.
- $A^{(\ell,h)} \in \mathbb{R}^{T \times T}$ for $h = 1, \ldots, H$: The attention matrices for each head. Entry $A_{ij}^{(\ell,h)}$ is the attention weight from token i to token j.

6.3.2 Step 1: Attention Graph Symmetrization

For each head h, compute the symmetrized attention matrix:

$$\tilde{A}^{(\ell,h)} = \frac{A^{(\ell,h)} + (A^{(\ell,h)})^\top}{2} \tag{31}$$

And the corresponding degree matrix and Laplacian:

$$\tilde{D}_{ii}^{(\ell,h)} = \sum_j \tilde{A}_{ij}^{(\ell,h)}, \qquad \tilde{L}^{(\ell,h)} = \tilde{D}^{(\ell,h)} - \tilde{A}^{(\ell,h)} \tag{32}$$

6.3.3 Step 2: Per-Head Energy Computation

For each head h, compute the Dirichlet energy:

$$E_D^{(\ell,h)} = \mathrm{Tr}\left((X^{(\ell)})^\top \tilde{L}^{(\ell,h)} X^{(\ell)} \right) = \sum_{i,j} \tilde{A}_{ij}^{(\ell,h)} \|x_i^{(\ell)} - x_j^{(\ell)}\|^2 \tag{33}$$

The pairwise form on the right avoids explicit Laplacian construction, requiring only the attention weights and hidden states.

6.3.4 Step 3: Head Aggregation

The layer energy is the mean across heads:

$$E_{\text{Attn}}^{(\ell)} = \frac{1}{H} \sum_{h=1}^{H} E_D^{(\ell,h)} \tag{34}$$

The mean is preferred over the sum for cross-architecture comparability: models with different head counts produce energy values on the same scale.

6.3.5 Step 4: Normalization

To enable comparison across inputs of different sequence lengths, the energy is normalized by the sequence length:

$$\hat{E}_{\text{Attn}}^{(\ell)} = \frac{E_{\text{Attn}}^{(\ell)}}{T} \tag{35}$$

This per-token normalization ensures that the coherence band thresholds are independent of sequence length. The invariant (Equation 25) is evaluated on the normalized energy $\hat{E}_{\text{Attn}}^{(\ell)}$, and the band parameters $E_{\min}^{D}(\ell)$, $E_{\max}^{D}(\ell)$ are calibrated on normalized values.

6.4 The DE Ratio Layer-Wise Diagnostic

6.4.1 Definition

The Dirichlet Energy Ratio at layer ℓ is:

$$R_{DE}^{(\ell)} = \frac{\hat{E}_{\text{Attn}}^{(\ell)}}{\hat{E}_{\text{Attn}}^{(\ell-1)}} \tag{36}$$

For the first monitored layer, the denominator uses the energy of the input embedding layer computed on a kNN graph (Section 3.3.4).

6.4.2 The Cumulative DE Product

$$\Pi_{DE}^{(\ell)} = \prod_{k=1}^{\ell} R_{DE}^{(k)} = \frac{\hat{E}_{\text{Attn}}^{(\ell)}}{\hat{E}_{\text{Attn}}^{(0)}} \tag{37}$$

6.4.3 DE Ratio Thresholds

The DE Ratio is a diagnostic, not a primary invariant. It does not trigger compliance state transitions directly but provides early warning:

Condition	Interpretation	Action
$R_{DE}^{(\ell)} < R_{\text{smooth}}$ for ≥ 3 consecutive layers	Progressive smoothing	Warning: collapse risk
$R_{DE}^{(\ell)} > R_{\text{sharp}}$ for ≥ 2 consecutive layers	Progressive sharpening	Warning: fragmentation risk
$\Pi_{DE}^{(L-1)} < \Pi_{\min}$	Cumulative collapse	Warning: energy floor approach
$\Pi_{DE}^{(L-1)} > \Pi_{\max}$	Cumulative sharpening	Warning: energy ceiling approach

The thresholds R_{smooth}, R_{sharp}, $\Pi_{\min}$, $\Pi_{\max}$ are calibrated from the validation distribution (Section 8).

6.5 The High-Frequency Energy Ratio

6.5.1 Practical Definition

The full HFER (Equation 5) requires eigendecomposition of the Laplacian, which is computationally expensive. For production monitoring, AI-7 specifies a practical approximation based on the attention-weighted energy decomposition.

Define the **local HFER** at layer ℓ, head h as follows. Partition the token pairs into "low-frequency" (high attention weight, similar representations) and "high-frequency" (low attention weight, dissimilar representations) based on the median attention weight $\tilde{A}^{(\ell,h)}_{\text{med}}$:

$$\text{HFER}^{(\ell)} = \frac{\sum_h \sum_{(i,j):\tilde{A}^{(\ell,h)}_{ij} < \tilde{A}^{(\ell,h)}_{\text{med}}} \tilde{A}^{(\ell,h)}_{ij} \|x_i^{(\ell)} - x_j^{(\ell)}\|^2}{\sum_h \sum_{i,j} \tilde{A}^{(\ell,h)}_{ij} \|x_i^{(\ell)} - x_j^{(\ell)}\|^2} \tag{38}$$

This approximation captures the fraction of total energy contributed by weakly-attended token pairs. In a healthy model, weakly-attended pairs should contribute *less* energy (they are weakly connected precisely because they are dissimilar and the model has learned to ignore them). In a fragmented model, weakly-attended pairs contribute disproportionate energy because the representations have become chaotic—even tokens the model does not attend to have wildly different representations.

6.5.2 HFER Threshold

HFER Diagnostic Threshold

A sustained $\text{HFER}^{(\ell)} > 0.5$ at any monitored layer triggers enhanced monitoring and contributes to the fragmentation risk assessment. An HFER exceeding 0.5 indicates that more than half the representation energy is in the "high-frequency" (weakly-attended) component—the representation is dominated by noise rather than structure.

6.6 Monitoring Frequency and Measurement Windows

6.6.1 Monitoring Frequency

The monitoring frequency is deployment-context-dependent:

Deployment Context	Monitoring Rate	Full Spectral Analysis	Rationale
Safety-critical (medical, defense)	Every request	Every 100 requests	Immediate detection required
High-risk (financial, legal)	10% of requests	Every 1,000 requests	Balanced detection and overhead
Standard (enterprise)	1% of requests	Every 10,000 requests	Amortized overhead
Low-risk (consumer)	0.1% of requests	Periodic (hourly/daily)	Minimal overhead

Table 4: Monitoring frequency by deployment context. "Monitoring Rate" refers to the per-request Dirichlet energy computation. "Full Spectral Analysis" includes Fiedler value computation, HFER, and effective rank.

6.6.2 Measurement Windows

Following the multi-scale windowing methodology established in Clause AI-6, the coherence metrics are aggregated over three time scales:

- **Per-request window:** The energy, rank, and Fiedler value for a single inference request. Provides immediate anomaly detection (e.g., adversarial input detection).
- **Session window** (W_{session}, default: 100 requests): The rolling statistics (mean, variance, percentiles) of energy over a session. Provides trend detection (e.g., progressive collapse during a continuous interaction).
- **Global window** (W_{global}, default: 10,000 requests): The population-level statistics of energy over the deployment. Provides drift detection (e.g., gradual representational degradation over weeks of operation).

6.7 Compliance State Machine

The structural coherence invariant uses a deterministic state machine with four states, mirroring the architecture established in Clause AI-6.

6.7.1 States

- **GREEN (Conforming):** All three invariants (energy, rank, connectivity) satisfied at all monitored layers. No DE Ratio or HFER warnings active.
- **AMBER (Warning):** Primary invariants satisfied, but one or more diagnostic warnings active: DE Ratio trending toward floor/ceiling, HFER elevated, or energy within a warning margin of the band boundary.
- **RED (Non-Conforming):** One or more invariants violated at one or more monitored layers. The model is operating in a certifiably pathological geometric regime.
- **UNKNOWN:** Insufficient measurement data to determine compliance (e.g., monitoring pipeline not yet initialized, or measurement failure).

6.7.2 Warning Thresholds

The warning state is triggered when the energy approaches the band boundary. Define the warning margins:

$$E^D_{\text{warn-low}}(\ell) = E^D_{\min}(\ell) + \alpha_w \cdot \left(E^D_{\text{median}}(\ell) - E^D_{\min}(\ell)\right) \tag{39}$$

$$E^D_{\text{warn-high}}(\ell) = E^D_{\max}(\ell) - \alpha_w \cdot \left(E^D_{\max}(\ell) - E^D_{\text{median}}(\ell)\right) \tag{40}$$

where $\alpha_w \in (0, 0.5)$ is the warning margin fraction (default: $\alpha_w = 0.2$) and $E^D_{\text{median}}(\ell)$ is the median energy observed during calibration. The warning zone occupies the outer 20% of each side of the band.

6.7.3 State Transition Rules

State Transition Rules

The following six rules govern deterministic state transitions:

1. **GREEN → AMBER:** Energy enters the warning zone ($\hat{E}^{(\ell)}_{\text{Attn}} < E^{D}_{\text{warn-low}}(\ell)$ or $\hat{E}^{(\ell)}_{\text{Attn}} > E^{D}_{\text{warn-high}}(\ell)$) at any monitored layer; OR a DE Ratio or HFER warning is triggered.

2. **AMBER → RED:** Energy exits the certified band ($\hat{E}^{(\ell)}_{\text{Attn}} < E^{D}_{\min}(\ell)$ or $\hat{E}^{(\ell)}_{\text{Attn}} > E^{D}_{\max}(\ell)$) at any monitored layer; OR effective rank drops below $r_{\min}(\ell)$; OR Fiedler value drops below $\tau_{\text{connected}}$.

3. **GREEN → RED:** Direct transition when a primary invariant is violated without passing through the warning zone (e.g., sudden adversarial spike).

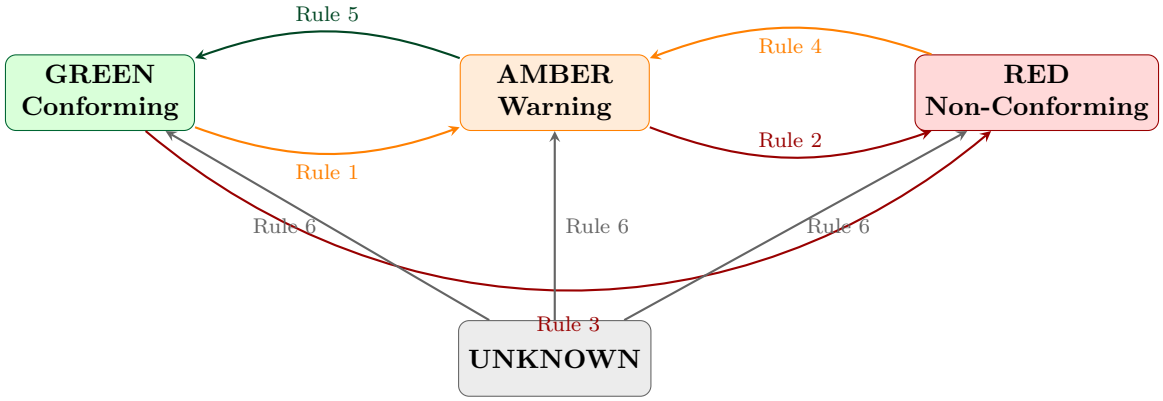
4. **RED → AMBER:** Energy returns to within the certified band AND rank recovers above $r_{\min}$ AND Fiedler value recovers above $\tau_{\text{connected}}$, sustained for $\geq W_{\text{recovery}}$ consecutive measurement points (default: $W_{\text{recovery}} = 10$).

5. **AMBER → GREEN:** All diagnostic warnings clear AND energy exits the warning zone (returns to the interior of the band), sustained for $\geq W_{\text{clear}}$ consecutive measurement points (default: $W_{\text{clear}} = 5$).

6. **UNKNOWN → GREEN/AMBER/RED:** Upon accumulation of sufficient measurement data ($\geq W_{\text{init}}$ measurements, default: $W_{\text{init}} = 20$), transition to the state determined by the collected data.

The recovery hysteresis (W_{recovery}, W_{clear}) prevents rapid oscillation between states due to measurement noise. A model that enters RED must demonstrate sustained recovery before returning to AMBER, and sustained health before returning to GREEN.

6.7.4 State Machine Diagram



6.8 Audit-Ready Parameter Table

The following table enumerates every parameter in the Clause AI-7 specification with its default value, calibration source, and governance rationale.

Parameter	Default	Calibration Source	Rationale
$E^D_{\min}(\ell)$	Model-specific	1st percentile of validation energy	Collapse floor: energy below this level is pathological
$E^D_{\max}(\ell)$	Model-specific	99th percentile of validation energy	Fragmentation ceiling: energy above this level is pathological
$r_{\min}(\ell)$	Model-specific	1st percentile of validation effective rank	Dimensional collapse floor
$\tau_{\text{connected}}$	Model-specific	5th percentile of validation Fiedler value	Connectivity threshold
α_w	0.2	Fixed	Warning margin fraction (outer 20% of band)
$\lvert\mathcal{M}\rvert$	4 layers	$\{L/4, L/2, 3L/4, L-1\}$	Monitoring coverage
R_{smooth}	0.85	5th percentile of validation DE Ratios	Smoothing warning threshold
R_{sharp}	1.20	95th percentile of validation DE Ratios	Sharpening warning threshold
$\Pi_{\min}$	0.3	1st percentile of validation cumulative DE	Cumulative collapse threshold
$\Pi_{\max}$	3.0	99th percentile of validation cumulative DE	Cumulative sharpening threshold
HFER threshold	0.5	Theoretical (energy equipartition)	High-frequency dominance indicator
W_{recovery}	10	Fixed	Measurement points for RED $\rightarrow$ AMBER
W_{clear}	5	Fixed	Measurement points for AMBER $\rightarrow$ GREEN
W_{init}	20	Fixed	Measurement points for UNKNOWN exit
Monitoring rate	Context-dependent	Table 4	Request sampling fraction

Table 5: Complete audit-ready parameter table for Clause AI-7. Model-specific parameters are calibrated from the validation distribution (Section 8). Fixed parameters are governance constants.

6.9 MAI-1 Layer 2 Payload Fields

The structural coherence measurements are carried in the MAI-1 Layer 2 attestation payload. The following fields are defined:

Field	Type	Description
`coherence-energy`	`float64[]`	Array of normalized Dirichlet energies $\hat{E}^{(\ell)}_{\text{Attn}}$ for each monitored layer $\ell \in \mathcal{M}$, in layer order.
`thresholds.coherence-energy-min`	`float64[]`	Certified collapse floor $E^D_{\min}(\ell)$ for each monitored layer.
`thresholds.coherence-energy-max`	`float64[]`	Certified fragmentation ceiling $E^D_{\max}(\ell)$ for each monitored layer.
`coherence-rank`	`float64[]`	Effective rank $\text{erank}(X^{(\ell)})$ for each monitored layer.
`thresholds.coherence-rank-min`	`float64[]`	Certified rank floor $r_{\min}(\ell)$ for each monitored layer.
`coherence-fiedler`	`float64[]`	Head-averaged Fiedler value $\bar{\lambda}^{(\ell)}_2$ for each monitored layer.
`coherence-hfer`	`float64[]`	High-Frequency Energy Ratio for each monitored layer.
`coherence-de-ratio`	`float64[]`	DE Ratio $R^{(\ell)}_{DE}$ for each monitored layer.
`coherence-state`	`enum`	Current compliance state: `GREEN`, `AMBER`, `RED`, or `UNKNOWN`.
`coherence-monitored-layers`	`uint8[]`	Layer indices comprising $\mathcal{M}$.
`coherence-timestamp`	`datetime`	Measurement timestamp (UTC).

Table 6: MAI-1 Layer 2 payload fields for Clause AI-7. All array fields are ordered by the monitored layer set $\mathcal{M}$.

6.10 Relationship to CTS-1 Conformance Assertions

The Clause AI-7 invariant generates the following testable assertions for the CTS-1 Conformance Test Suite:

Assertion ID	Description
`AS-SM-L2-20`	Energy band: $E^D_{\min}(\ell) \leq \hat{E}^{(\ell)}_{\text{Attn}} \leq E^D_{\max}(\ell)$
`AS-SM-L2-20.01`	Rank floor: $\text{erank}(X^{(\ell)}) \geq r_{\min}(\ell)$
`AS-SM-L2-20.02`	Fiedler connectivity: $\bar{\lambda}^{(\ell)}_2 \geq \tau_{\text{connected}}$
`TE-SM-L2-20.01`	State machine: transitions follow Rules 1–6
`TE-SM-L2-20.02`	Recovery hysteresis: W_{recovery} sustained points
`TE-SM-L2-20.03`	Layer coverage: $\lvert\mathcal{M}\rvert \geq 4$
`TE-SM-L2-20.04`	Payload completeness: all Table 6 fields present

Each assertion has a binary verdict: PASS or FAIL. There is no partial credit. A system claiming Clause AI-7 conformance **SHALL** pass all assertions.

7 Transformer-Specific Measurement Methodology

Section 6 defined *what* to measure. This section specifies *how* to measure it—the exact algorithms, data flows, and implementation decisions that transform the mathematical definitions into executable monitoring procedures for production transformer architectures. This is the methodological core is the bridge between geometric theory and deployed infrastructure.

7.1 Architecture Assumptions

The measurement methodology is specified for the dominant production architecture: the decoder-only transformer with multi-head self-attention, pre-layer normalization, rotary position embeddings, and grouped-query attention (GQA). The methodology generalizes to encoder-only and encoder-decoder architectures with minor modifications noted inline.

Architecture Interface Requirements

A model claiming Clause AI-7 conformance **MUST** expose the following intermediate values at each monitored layer $\ell \in \mathcal{M}$:

1. The post-attention hidden state matrix $X^{(\ell)} \in \mathbb{R}^{T \times d}$, captured *after* the attention sublayer and its residual connection but *before* the feed-forward sublayer.
2. The attention weight matrices $A^{(\ell,h)} \in \mathbb{R}^{T \times T}$ for each attention head $h = 1, \ldots, H$, captured *after* softmax normalization.

For models using Grouped-Query Attention (GQA), where H_q query heads share H_{kv} key-value heads ($H_q > H_{kv}$), the attention matrices **SHALL** be reported per *query head* (i.e., $H = H_q$ matrices), not per key-value group. Each query head produces a distinct attention pattern even when sharing key-value projections.

For models using Multi-Query Attention (MQA, $H_{kv} = 1$), each of the H_q query heads still produces a unique attention matrix. The measurement methodology applies without modification.

7.2 The Attention-Weighted Dirichlet Energy Algorithm

7.2.1 Algorithm Specification

The following algorithm computes the normalized attention-weighted Dirichlet energy $\hat{E}^{(\ell)}_{\text{Attn}}$ at a single monitored layer. It is designed for GPU execution with standard BLAS-3 operations.

Algorithm 1: Attention-Weighted Dirichlet Energy

Input: Hidden states $X^{(\ell)} \in \mathbb{R}^{T \times d}$, attention matrices $\{A^{(\ell,h)}\}_{h=1}^{H}$
Output: Normalized energy $\hat{E}_{\text{Attn}}^{(\ell)}$

1. **Compute pairwise squared distances.**
Construct the squared distance matrix $\Delta \in \mathbb{R}^{T \times T}$:

$$\Delta_{ij} = \|x_i^{(\ell)} - x_j^{(\ell)}\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2\, x_i^\top x_j \tag{41}$$

Implementation: compute $\mathbf{n} = \text{diag}(XX^\top) \in \mathbb{R}^T$ (row norms squared), the Gram matrix $G = XX^\top \in \mathbb{R}^{T \times T}$, and assemble $\Delta = \mathbf{n}\mathbf{1}^\top + \mathbf{1}\mathbf{n}^\top - 2G$.
Complexity: $O(T^2 d)$ for the Gram matrix (one GEMM call).

2. **Symmetrize attention matrices.**
For each head h:

$$\tilde{A}^{(\ell,h)} = \frac{A^{(\ell,h)} + (A^{(\ell,h)})^\top}{2} \tag{42}$$

Complexity: $O(H \cdot T^2)$.

3. **Compute per-head energy via Hadamard product.**
For each head h:

$$E_D^{(\ell,h)} = \sum_{i,j} \tilde{A}_{ij}^{(\ell,h)} \cdot \Delta_{ij} = \langle \tilde{A}^{(\ell,h)}, \Delta \rangle_F \tag{43}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product (element-wise multiply and sum).
Complexity: $O(T^2)$ per head (Hadamard product + reduction).

4. **Aggregate across heads.**

$$E_{\text{Attn}}^{(\ell)} = \frac{1}{H} \sum_{h=1}^{H} E_D^{(\ell,h)} \tag{44}$$

5. **Normalize by sequence length.**

$$\hat{E}_{\text{Attn}}^{(\ell)} = \frac{E_{\text{Attn}}^{(\ell)}}{T} \tag{45}$$

Total complexity: $O(T^2 d + H \cdot T^2) = O(T^2(d + H))$ per monitored layer.
Memory: $O(T^2)$ for Δ (shared across heads) plus $O(T^2)$ for one attention matrix at a time (heads can be processed sequentially to avoid storing all H symmetrized matrices simultaneously).

7.2.2 Numerical Considerations

Precision. The pairwise distance computation $\Delta_{ij} = \|x_i\|^2 + \|x_j\|^2 - 2x_i^\top x_j$ is susceptible to catastrophic cancellation when $x_i \approx x_j$ (which is precisely the regime of interest for collapse detection). Implementations **SHALL** compute Δ in FP32 even if the hidden states are stored in FP16/BF16. The Gram matrix $G = XX^\top$ **SHALL** use FP32 accumulation.

Causal masking. In decoder-only transformers, the attention matrix is lower-triangular (token i can only attend to tokens $j \leq i$). The symmetrization $\tilde{A} = (A + A^\top)/2$ creates a symmetric matrix from this triangular structure. Entries $\tilde{A}_{ij}$ for $i < j$ become $A_{ji}/2$ (the attention that token j pays to token i, halved). This is the correct behavior: the symmetrized graph captures mutual relevance between tokens, regardless of the autoregressive ordering.

Zero attention entries. For very long sequences with sparse attention patterns (e.g., sliding window attention), many entries of $A^{(\ell,h)}$ are exactly zero. The Hadamard product $\tilde{A} \odot \Delta$ inherits this sparsity, and the reduction can exploit it for computational savings. See Section 9 for the Top-K sparsification that formalizes this.

7.3 Head Aggregation Strategy

7.3.1 Mean Aggregation (Default)

The default head aggregation (Equation 34) takes the arithmetic mean across heads. This is appropriate when the diagnostic goal is to measure *average* representational coherence across the model's attention mechanisms.

7.3.2 Per-Head Diagnostics

For enhanced monitoring (safety-critical deployments), per-head energy values provide additional diagnostic power:

- **Head energy variance:** $\mathrm{Var}_h(E_D^{(\ell,h)})$ measures the dispersion of energy across heads. High variance indicates head specialization—some heads are smoothing while others are sharpening. This is normal in healthy models. A sudden change in head energy variance (compared to the validation baseline) signals a structural shift.
- **Outlier head detection:** A head with $E_D^{(\ell,h)} > \mu_h + 3\sigma_h$ (where μ_h, σ_h are the head's baseline mean and standard deviation) may indicate a head-specific pathology—for example, an attention head that has "died" (attending uniformly, producing zero energy) or "exploded" (attending chaotically, producing extreme energy).
- **Head-specific collapse:** If one or more heads consistently produce near-zero energy while others remain healthy, this indicates partial collapse—a failure mode where the model loses some of its attention diversity without complete representational breakdown.

7.3.3 GQA-Aware Aggregation

In Grouped-Query Attention architectures, query heads within the same key-value group share key-value projections but differ in their query projections. The energy values of heads within a group are correlated. For GQA models, AI-7 recommends a two-level aggregation:

$$E_{\text{Attn}}^{(\ell)} = \frac{1}{H_{kv}} \sum_{g=1}^{H_{kv}} \left(\frac{1}{|G_g|} \sum_{h \in G_g} E_D^{(\ell,h)} \right) \tag{46}$$

where G_g is the set of query heads in key-value group g. This first averages within groups (capturing intra-group variation) and then averages across groups (capturing inter-group variation), preventing groups with more query heads from dominating the aggregate.

7.4 Effective Rank Computation

7.4.1 Full SVD Method

The exact effective rank (Equation 13) requires the singular value decomposition of $X^{(\ell)} \in \mathbb{R}^{T \times d}$:

$$X^{(\ell)} = U \Sigma V^\top \tag{47}$$

The singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(T,d)}$ are normalized and the Shannon entropy computed:

$$\tilde{\sigma}_i = \frac{\sigma_i}{\sum_j \sigma_j}, \qquad \text{erank}(X^{(\ell)}) = \exp\left(-\sum_i \tilde{\sigma}_i \log \tilde{\sigma}_i\right) \tag{48}$$

Full SVD has complexity $O(\min(T, d) \cdot T \cdot d)$, which is expensive for large models. Section 9 specifies the randomized SVD approximation that reduces this to $O(T \cdot d \cdot \log k + (T + d) \cdot k^2)$ for target rank k.

7.4.2 Gram Matrix Method

When $T < d$ (typical for moderate sequence lengths with large hidden dimensions), the effective rank can be computed more efficiently via the Gram matrix:

$$K = X^{(\ell)}(X^{(\ell)})^\top \in \mathbb{R}^{T \times T} \tag{49}$$

The eigenvalues of K are the squared singular values of $X^{(\ell)}$: $\mu_i = \sigma_i^2$. The effective rank is:

$$\text{erank}(X^{(\ell)}) = \exp\left(-\sum_i \tilde{\mu}_i \log \tilde{\mu}_i\right), \qquad \tilde{\mu}_i = \frac{\sqrt{\mu_i}}{\sum_j \sqrt{\mu_j}} \tag{50}$$

Gram matrix eigendecomposition has complexity $O(T^2 d + T^3)$, which is cheaper than full SVD when $T \ll d$. For a 7B model with $d = 4096$ and $T = 2048$, this saves approximately 50% computation compared to full SVD.

7.4.3 Centering

The hidden state matrix **SHALL** be mean-centered before computing the effective rank:

$$\bar{X}^{(\ell)} = X^{(\ell)} - \frac{1}{T}\mathbf{1}\mathbf{1}^\top X^{(\ell)} \tag{51}$$

Centering removes the contribution of the mean representation (which inflates the largest singular value) and ensures that the effective rank measures the *diversity* of representations around their centroid, not the magnitude of the centroid itself. Without centering, a representation where all tokens are nearly identical but with a large mean norm would appear to have high rank due to the dominant first singular value encoding the mean.

7.5 Fiedler Value Computation

7.5.1 Exact Method

The Fiedler value is the second-smallest eigenvalue of the symmetrized attention Laplacian $\tilde{L}^{(\ell,h)}$. The head-averaged Fiedler value is:

$$\bar{\lambda}_2^{(\ell)} = \frac{1}{H}\sum_{h=1}^{H} \lambda_2(\tilde{L}^{(\ell,h)}) \tag{52}$$

Computing λ_2 for a single head requires a partial eigendecomposition of the $T \times T$ Laplacian. The smallest eigenvalue $\lambda_1 = 0$ is known (the constant eigenvector), so λ_2 can be found by computing the two smallest eigenvalues and taking the second.

7.5.2 Inverse Power Iteration

For production monitoring, the Fiedler value can be approximated efficiently using inverse power iteration on the shifted Laplacian. Since $\lambda_1 = 0$, we project out the constant eigenvector and apply power iteration to find the smallest eigenvalue of the projected Laplacian:

1. Initialize random vector $v_0 \in \mathbb{R}^T$ with $\mathbf{1}^\top v_0 = 0$ (orthogonal to the constant eigenvector).
2. For $k = 1, \ldots, M$ iterations:
   (a) Solve $\tilde{L}w = v_{k-1}$ (linear system).
   (b) Project: $w \leftarrow w - (\mathbf{1}^\top w/T) \cdot \mathbf{1}$.
   (c) Normalize: $v_k = w/\|w\|$.
   (d) Estimate: $\hat{\lambda}_2 = v_k^\top \tilde{L} v_k / (v_k^\top v_k)$ (Rayleigh quotient).

Convergence rate is $|\lambda_2/\lambda_3|$ per iteration. For well-separated eigenvalues (typical in attention graphs), 10–20 iterations suffice.

7.5.3 Stochastic Estimation

For the highest-throughput monitoring, the Fiedler value can be bounded rather than computed exactly. The Cheeger inequality (Equation 7) provides:

$$\lambda_2 \geq \frac{h^2}{2} \tag{53}$$

The Cheeger constant h can be approximated from the attention matrix by finding the minimum normalized cut. If the approximated h exceeds $\sqrt{2\tau_{\text{connected}}}$, then $\lambda_2 \geq \tau_{\text{connected}}$ is guaranteed without computing λ_2 directly. This provides a fast "pass" determination; only when the Cheeger bound is inconclusive does the full Fiedler computation trigger.

7.6 HFER Computation

7.6.1 The Practical HFER Algorithm

The practical HFER (Equation 38) avoids eigendecomposition by partitioning energy contributions based on attention weight magnitude:

Algorithm 2: Practical HFER

Input: Symmetrized attention matrices $\{\tilde{A}^{(\ell,h)}\}_{h=1}^{H}$, squared distance matrix Δ
Output: $\text{HFER}^{(\ell)}$

1. Compute the element-wise energy contribution matrix:

$$C^{(\ell,h)} = \tilde{A}^{(\ell,h)} \odot \Delta \tag{54}$$

2. For each head h, compute the median attention weight:

$$\tilde{A}^{(\ell,h)}_{\text{med}} = \text{median}\left(\{\tilde{A}^{(\ell,h)}_{ij} : \tilde{A}^{(\ell,h)}_{ij} > 0\}\right) \tag{55}$$

3. Compute the high-frequency energy (contributions from below-median attention):

$$E^{(\ell)}_{\text{HF}} = \sum_{h} \sum_{(i,j):\tilde{A}^{(\ell,h)}_{ij} < \tilde{A}^{(\ell,h)}_{\text{med}}} C^{(\ell,h)}_{ij} \tag{56}$$

4. Compute total energy:

$$E^{(\ell)}_{\text{total}} = \sum_{h} \sum_{i,j} C^{(\ell,h)}_{ij} \tag{57}$$

5. Return: $\text{HFER}^{(\ell)} = E^{(\ell)}_{\text{HF}} / E^{(\ell)}_{\text{total}}$.

Complexity: $O(H \cdot T^2)$ (dominated by the median computation and partitioned summation, which reuse the already-computed Δ from Algorithm 1).

7.6.2 Interpretation of the Practical HFER

The practical HFER partitions energy by the attention weight threshold rather than by Laplacian eigenvalue index. The correspondence is approximate but informative:

- High attention weight $\tilde{A}_{ij}$ indicates a "strong edge" in the attention graph. Energy from strong edges corresponds to low-frequency variation—the model actively connects these tokens and their representation difference is "expected" variation.
- Low attention weight indicates a "weak edge." Energy from weak edges corresponds to high-frequency variation—these tokens are weakly connected but have large representation differences, indicating noise, fragmentation, or adversarial perturbation.

A healthy model concentrates its energy on strong edges (HFER < 0.5). A fragmented model distributes energy across weak edges (HFER > 0.5).

7.7 The DE Ratio Computation

7.7.1 Standard DE Ratio

The DE Ratio (Equation 36) is computed as the ratio of normalized energies at consecutive monitored layers. When the monitored layers are not consecutive (e.g., $\mathcal{M} = \{8, 16, 24, 31\}$ for a 32-layer model), the DE Ratio between monitored layers ℓ_k and ℓ_{k+1} captures the *cumulative* smoothing/sharpening effect of all intervening layers:

$$R_{DE}^{(\ell_{k+1})} = \frac{\hat{E}_{\text{Attn}}^{(\ell_{k+1})}}{\hat{E}_{\text{Attn}}^{(\ell_k)}} \tag{58}$$

This inter-checkpoint ratio is the production-relevant quantity: it measures the aggregate energy change between monitored layers without requiring energy computation at every intermediate layer.

7.7.2 Initial Energy Baseline

The DE Ratio for the first monitored layer requires a reference energy at the input. Since no attention matrix exists at layer 0, the initial energy is computed using a k-nearest-neighbor graph on the input embeddings:

1. Let $X^{(0)} \in \mathbb{R}^{T \times d}$ be the token embedding matrix (after positional encoding).
2. Compute pairwise cosine similarities: $S_{ij} = x_i^\top x_j / (\|x_i\| \|x_j\|)$.
3. Construct the kNN graph: for each token i, connect to its k_0 nearest neighbors (default $k_0 = 16$).
4. Symmetrize: $\tilde{A}_{ij}^{(0)} = \max(A_{ij}^{(0)}, A_{ji}^{(0)})$.
5. Compute the Dirichlet energy on this graph using the same algorithm as for attention layers.

The kNN parameter k_0 is calibrated during the validation phase to produce an initial energy in the same range as the first monitored layer. The sensitivity of the DE Ratio to k_0 is characterized during calibration (Section 8).

7.8 Measurement Point Selection for Causal Models

7.8.1 The Sequence Length Challenge

In autoregressive (causal) language models, the hidden states evolve as each token is generated. The representation $X^{(\ell)}$ at layer ℓ changes with every generated token because the attention matrix incorporates the new token. This raises the question: at which point in the generation process should the coherence measurement be taken?

7.8.2 Specification

Measurement Point for Autoregressive Models

For causal language models, the structural coherence measurement **SHALL** be taken at the following points:

1. **Prompt completion:** After the full prompt (input context) has been processed, before generation begins. This measures the coherence of the model's representation of the input—a necessary precondition for coherent generation.
2. **Generation checkpoints:** At intervals of Δ_{gen} generated tokens (default: $\Delta_{\text{gen}} = 64$). This detects coherence degradation *during* generation—the inertial drift and attention fragmentation pathologies described in Section 2.
3. **Generation completion:** After the final token is generated. This provides the terminal coherence state for the attestation record.

The per-token energy can be computed incrementally (Section 9.3) to avoid recomputing the full $T \times T$ distance matrix at each checkpoint.

7.8.3 The Growing Sequence Problem

As generation proceeds, the sequence length T grows. The normalized energy $\hat{E}^{(\ell)}_{\text{Attn}} = E^{(\ell)}_{\text{Attn}}/T$ accounts for this growth, ensuring that the coherence band thresholds remain applicable across different sequence lengths. However, the computational cost of each measurement grows as $O(T^2)$, which motivates the Top-K sparsification and sampling strategies in Section 9.

7.9 Multi-Batch Aggregation

7.9.1 Batch-Level Statistics

In high-throughput serving systems, multiple requests are batched together. The coherence measurement is computed *per request* (each request has its own attention matrices and hidden states), but batch-level aggregation provides population-level diagnostics:

- **Batch mean energy:** $\bar{E}^{(\ell)}_{\text{batch}} = \frac{1}{B}\sum_{b=1}^{B} \hat{E}^{(\ell)}_{\text{Attn},b}$. Tracks the population-level energy trend.
- **Batch energy variance:** $\text{Var}^{(\ell)}_{\text{batch}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{E}^{(\ell)}_{\text{Attn},b} - \bar{E}^{(\ell)}_{\text{batch}})^2$. A sudden increase in variance (compared to the validation baseline) indicates that the model is producing inconsistent representational geometry across inputs—a potential sign of distributional shift or mode collapse.
- **Batch outlier count:** The number of requests in the batch with energy outside the coherence band. A high outlier rate triggers enhanced monitoring.

7.9.2 Rolling Statistics

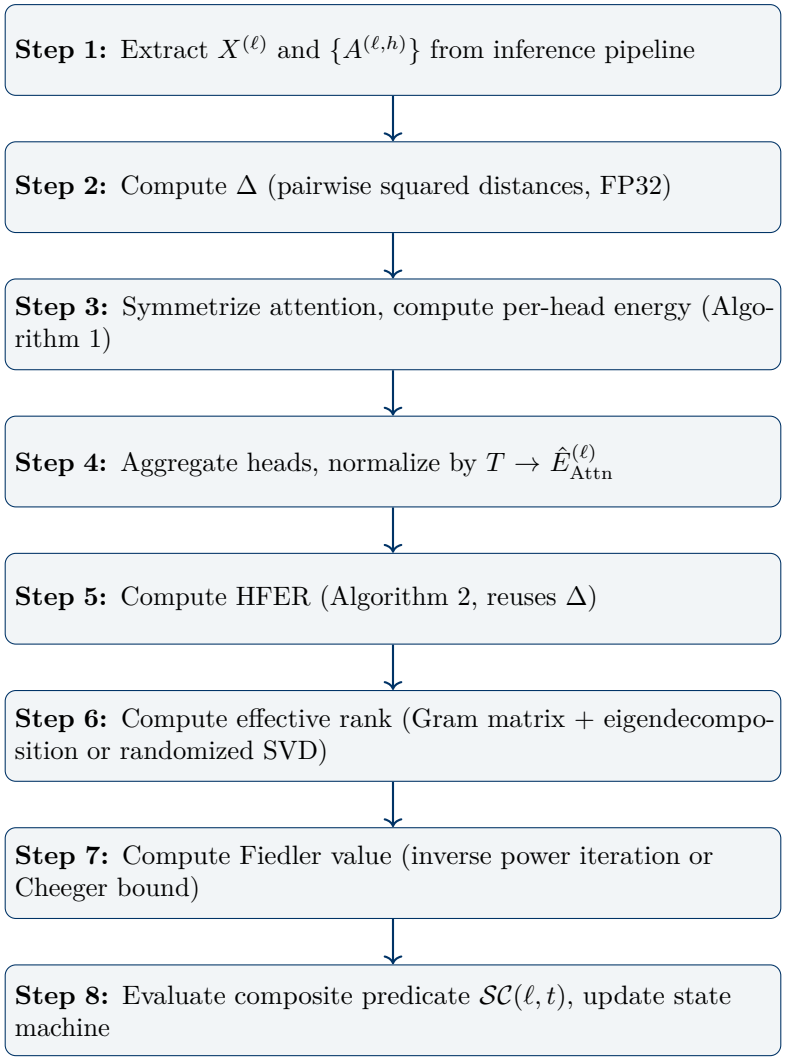
Session-level and global-level statistics (Section 6.5) are maintained via exponentially weighted moving averages:

$$\bar{E}^{(\ell)}_{\text{session}}(t) = \beta \cdot \bar{E}^{(\ell)}_{\text{session}}(t-1) + (1-\beta) \cdot \hat{E}^{(\ell)}_{\text{Attn}}(t) \tag{59}$$

where $\beta \in (0,1)$ is the smoothing factor (default: $\beta = 0.99$ for session-level, $\beta = 0.999$ for global-level). The EWMA provides efficient constant-memory tracking of the energy trend without storing individual measurements.

7.10 Summary: The Measurement Pipeline

The complete measurement pipeline for a single monitored layer at a single measurement point is:



Steps 2–5 are computationally coupled (they share the Δ matrix). Steps 6–7 are independent and can execute in parallel. The total pipeline is repeated for each monitored layer $\ell \in \mathcal{M}$ (4 layers by default). Section 9 specifies the optimizations that make this pipeline feasible within the $< 2\%$ latency budget.

8 Calibration Procedure for the Certified Coherence Band

The coherence band $[E^D_{\min}(\ell), E^D_{\max}(\ell)]$, the effective rank floor $r_{\min}(\ell)$, and the Fiedler connectivity threshold $\tau_{\text{connected}}$ are model-specific parameters that **MUST** be derived from empirical measurement on a validation distribution. This section specifies the complete calibration procedure: the validation data requirements, the statistical methodology for establishing each threshold, the conformal coverage guarantee that links calibration to governance confidence, and the recalibration protocol for model updates.

8.1 Calibration Philosophy

The calibration procedure follows the same design philosophy established in Clause AI-6 for the distribution drift threshold:

1. **The model defines its own band.** The coherence band is not a universal constant; it is derived from the model's own energy profile on representative data. Different architectures,

training procedures, and task domains produce different energy landscapes. A 7B-parameter model and a 70B-parameter model may have entirely different healthy energy ranges.

2. **The calibration data must be representative.** The validation distribution used for calibration **SHALL** represent the intended deployment distribution. Calibrating on a narrow benchmark and deploying on a broad distribution will produce a band that is too tight (excessive false alarms). Calibrating on an excessively broad distribution will produce a band that is too loose (missed pathologies).

3. **The thresholds are conservative.** The floor and ceiling are set at extreme percentiles (1st and 99th) of the empirical distribution, ensuring that the band contains the vast majority of healthy operating points. The governance risk margin provides additional buffer for high-risk deployments.

4. **Conformal coverage provides a distribution-free guarantee.** The calibration uses conformal prediction methodology to provide a finite-sample coverage guarantee that does not depend on distributional assumptions about the energy values.

8.2 Validation Data Requirements

8.2.1 Dataset Specification

Calibration Dataset Requirements

The calibration dataset $\mathcal{D}_{\text{cal}}$ **SHALL** satisfy the following:

1. **Size:** $|\mathcal{D}_{\text{cal}}| \geq N_{\text{cal}}$, where $N_{\text{cal}} = 1{,}000$ is the minimum for governance-grade calibration. For conformal coverage at level $1 - \alpha$ with tolerance ϵ, the minimum sample size is $N_{\text{cal}} \geq \lceil (1 + 1/\epsilon)/\alpha \rceil$. At $\alpha = 0.01$ and $\epsilon = 0.001$: $N_{\text{cal}} \geq 1{,}001$.
2. **Representativeness:** The dataset **SHALL** be drawn from the intended deployment distribution or a close approximation thereof. If the deployment distribution is unknown, the calibration dataset **SHALL** cover the expected input diversity (language, domain, length distribution, task types).
3. **Clean data:** The dataset **SHALL** be free of known adversarial examples, poisoned data, or out-of-distribution anomalies. The purpose of calibration is to establish the healthy operating regime; inclusion of pathological inputs would inflate the band.
4. **Length diversity:** The dataset **SHALL** include inputs spanning the expected range of sequence lengths. The normalized energy $\hat{E}^{(\ell)}_{\text{Attn}} = E^{(\ell)}_{\text{Attn}}/T$ should be approximately length-independent for healthy inputs; if systematic length dependence is observed during calibration, the band **SHALL** be stratified by length bin (Section 8.4).
5. **Independence:** Calibration samples **SHALL** be independent. Sequential samples from the same document or conversation violate exchangeability and invalidate conformal coverage guarantees.

8.2.2 Calibration Measurements

For each sample $x_n \in \mathcal{D}_{\text{cal}}$, the measurement pipeline (Section 7.8) is executed at each monitored layer $\ell \in \mathcal{M}$, producing:

- $\hat{E}^{(\ell)}_{\text{Attn},n}$: Normalized Dirichlet energy at layer ℓ.
- $\text{erank}^{(\ell)}_n$: Effective rank at layer ℓ.
- $\bar{\lambda}^{(\ell)}_{2,n}$: Head-averaged Fiedler value at layer ℓ.

- $R_{DE,n}^{(\ell)}$: DE Ratio at layer ℓ.
- $\text{HFER}_n^{(\ell)}$: High-Frequency Energy Ratio at layer ℓ.

This produces, for each layer, N_{cal} measurements of each quantity.

8.3 Threshold Derivation: The Three-Step Methodology

The threshold derivation follows a three-step process: empirical percentile estimation, conformal calibration, and governance risk margin application.

8.3.1 Step 1: Empirical Percentile Estimation

For each monitored layer ℓ, sort the N_{cal} energy measurements:

$$\hat{E}_{(1)}^{(\ell)} \leq \hat{E}_{(2)}^{(\ell)} \leq \cdots \leq \hat{E}_{(N_{\text{cal}})}^{(\ell)} \tag{60}$$

Compute the empirical percentile thresholds:

$$\hat{E}_{\text{floor}}^{(\ell)} = \hat{E}_{(\lceil 0.01 \cdot N_{\text{cal}} \rceil)}^{(\ell)} \quad \text{(1st percentile)} \tag{61}$$

$$\hat{E}_{\text{ceiling}}^{(\ell)} = \hat{E}_{(\lceil 0.99 \cdot N_{\text{cal}} \rceil)}^{(\ell)} \quad \text{(99th percentile)} \tag{62}$$

$$\hat{E}_{\text{median}}^{(\ell)} = \hat{E}_{(\lceil 0.50 \cdot N_{\text{cal}} \rceil)}^{(\ell)} \quad \text{(median, for warning zone computation)} \tag{63}$$

Similarly, for the effective rank:

$$\hat{r}_{\text{floor}}^{(\ell)} = \text{erank}_{(\lceil 0.01 \cdot N_{\text{cal}} \rceil)}^{(\ell)} \quad \text{(1st percentile of rank distribution)} \tag{64}$$

And for the Fiedler value:

$$\hat{\tau}_{\text{floor}}^{(\ell)} = \bar{\lambda}_{2,(\lceil 0.05 \cdot N_{\text{cal}} \rceil)}^{(\ell)} \quad \text{(5th percentile of Fiedler distribution)} \tag{65}$$

The 5th percentile (rather than 1st) for the Fiedler value reflects the higher natural variability of connectivity measurements: attention graph topology is more input-sensitive than energy or rank.

8.3.2 Step 2: Conformal Calibration

The empirical percentiles from Step 1 provide point estimates. Conformal prediction provides a *finite-sample coverage guarantee* without distributional assumptions, following the methodology established in Clause AI-6.

Conformal coverage guarantee. Given N_{cal} exchangeable calibration samples, the conformal prediction framework guarantees that for a new test sample $x_{N_{\text{cal}}+1}$:

$$\mathbb{P}\left[E_{\min}^{D}(\ell) \leq \hat{E}_{\text{Attn},N_{\text{cal}}+1}^{(\ell)} \leq E_{\max}^{D}(\ell)\right] \geq 1 - \alpha \tag{66}$$

where α is the miscoverage rate. For $\alpha = 0.01$, the band contains at least 99% of future healthy observations.

Conformal floor and ceiling. The conformal adjustment expands the empirical band to account for finite-sample uncertainty:

$$E_{\min,\text{conf}}^{D}(\ell) = \hat{E}_{(\lfloor (N_{\text{cal}}+1)\cdot\alpha/2 \rfloor)}^{(\ell)} \tag{67}$$

$$E_{\max,\text{conf}}^{D}(\ell) = \hat{E}_{(\lceil (N_{\text{cal}}+1)\cdot(1-\alpha/2) \rceil)}^{(\ell)} \tag{68}$$

The $\alpha/2$ split allocates equal miscoverage probability to each tail. For $N_{\text{cal}} = 1{,}000$ and $\alpha = 0.01$: the conformal floor is the 5th order statistic and the conformal ceiling is the 996th order statistic.

The conformal guarantee (Equation 66) holds under a single assumption: *exchangeability* of the calibration and test samples. This is strictly weaker than the i.i.d. assumption. It does not require the energy values to follow any parametric distribution. This distribution-free property is essential for governance applications, where the energy distribution may be heavy-tailed, multimodal, or otherwise non-standard.

8.3.3 Step 3: Governance Risk Margin

The conformal band provides statistical coverage. The governance risk margin provides additional conservatism for high-risk deployment contexts, following the same risk-tier structure used in Clause AI-6:

$$E^D_{\min}(\ell) = E^D_{\min,\text{conf}}(\ell) \cdot (1 - \gamma_{\text{floor}}) \tag{69}$$

$$E^D_{\max}(\ell) = E^D_{\max,\text{conf}}(\ell) \cdot (1 + \gamma_{\text{ceiling}}) \tag{70}$$

where $\gamma_{\text{floor}}, \gamma_{\text{ceiling}} \in [0, 1)$ are the governance risk margins.

Deployment Context	γ_{floor}	γ_{ceiling}	**Rationale**
Safety-critical (medical, defense, autonomous systems)	0.15	0.15	Widest margin. False negatives (missed pathologies) are unacceptable.
High-risk (financial, legal, critical infrastructure)	0.10	0.10	Moderate margin. Regulatory scrutiny demands conservatism.
Standard (enterprise, customer-facing)	0.05	0.05	Narrow margin. Balances detection sensitivity with operational overhead.
Low-risk (internal tools, non-critical applications)	0.00	0.00	No margin. Conformal band only.

Table 7: Governance risk margins by deployment context. The margin widens the certified band, reducing false alarms (healthy inputs flagged as pathological) at the cost of slightly reduced sensitivity to genuine pathology. Higher-risk contexts use larger margins to ensure that the band is conservative even under worst-case calibration uncertainty.

Applying the risk margin to rank and Fiedler thresholds:

$$r_{\min}(\ell) = \hat{r}^{(\ell)}_{\text{floor}} \cdot (1 - \gamma_{\text{floor}}) \tag{71}$$

$$\tau_{\text{connected}} = \hat{\tau}_{\text{floor}} \cdot (1 - \gamma_{\text{floor}}) \tag{72}$$

The Fiedler threshold $\tau_{\text{connected}}$ is global (not layer-specific) because the minimum connectivity required for coherent processing is an architectural property, not a layer-specific one. It is computed as the minimum across layers of the per-layer 5th percentile, then adjusted by the governance margin.

8.4 Length Stratification

8.4.1 When Stratification Is Required

The normalized energy $\hat{E}^{(\ell)}_{\text{Attn}}$ should be approximately independent of sequence length T for a healthy model. However, systematic length dependence can arise from:

- Attention pattern changes: short sequences produce dense attention matrices (every token attends to every other token), while long sequences produce sparser effective attention (each token concentrates on a subset).
- Positional encoding effects: rotary position embeddings (RoPE) attenuate attention between distant tokens, altering the attention graph topology and therefore the energy.
- Padding and batching artifacts: padded sequences may produce anomalous energy if padding tokens are included in the computation.

8.4.2 Length Dependence Test

During calibration, test for systematic length dependence by computing the Spearman rank correlation ρ_S between sequence length and normalized energy:

$$\rho_S = \text{Spearman}(T_n, \hat{E}^{(\ell)}_{\text{Attn},n}) \quad \text{for } n = 1, \ldots, N_{\text{cal}} \tag{73}$$

If $|\rho_S| > 0.3$ at any monitored layer, length stratification **SHALL** be applied.

8.4.3 Stratified Calibration

When stratification is required, partition the calibration data into B_T length bins:

$$\mathcal{D}_{\text{cal}} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_{B_T} \tag{74}$$

where $\mathcal{D}_b = \{x_n : T_n \in [T_b^{\text{lo}}, T_b^{\text{hi}})\}$. Default bins: $[1, 256)$, $[256, 1024)$, $[1024, 4096)$, $[4096, \infty)$. Each bin receives its own coherence band:

$$[E^D_{\min}(\ell, b),\ E^D_{\max}(\ell, b)] \quad \text{for each layer } \ell \text{ and length bin } b \tag{75}$$

At inference time, the appropriate band is selected based on the input sequence length. The MAI-1 payload carries the active band parameters.

8.5 Calibration for DE Ratio Thresholds

The DE Ratio thresholds $(R_{\text{smooth}}, R_{\text{sharp}}, \Pi_{\min}, \Pi_{\max})$ are calibrated from the empirical distribution of DE Ratios observed during validation:

$$R_{\text{smooth}} = \text{Percentile}_5\left(\{R^{(\ell)}_{DE,n} : n = 1, \ldots, N_{\text{cal}}\}\right) \tag{76}$$

$$R_{\text{sharp}} = \text{Percentile}_{95}\left(\{R^{(\ell)}_{DE,n} : n = 1, \ldots, N_{\text{cal}}\}\right) \tag{77}$$

$$\Pi_{\min} = \text{Percentile}_1\left(\{\Pi^{(L-1)}_{DE,n} : n = 1, \ldots, N_{\text{cal}}\}\right) \tag{78}$$

$$\Pi_{\max} = \text{Percentile}_{99}\left(\{\Pi^{(L-1)}_{DE,n} : n = 1, \ldots, N_{\text{cal}}\}\right) \tag{79}$$

These are diagnostic thresholds (warning triggers, not compliance boundaries), so conformal adjustment and governance margins are not applied.

8.6 Band Width and Model Capacity

8.6.1 The Relationship Between Band Width and Architecture

The width of the coherence band $\Delta E^{(\ell)} = E^D_{\max}(\ell) - E^D_{\min}(\ell)$ is not a free parameter; it is an *emergent property* of the model's architecture, training procedure, and deployment distribution. Empirically, the band width correlates with:

- **Model depth:** Deeper models exhibit wider bands because more layers produce more diverse energy dynamics. The DE Ratio product has more terms, and the variance of the cumulative product increases with depth.
- **Attention head count:** More heads produce greater energy diversity (different heads specialize in different frequency ranges), widening the band.
- **Hidden dimension:** Larger hidden dimensions provide more "room" for representations to vary, increasing the range of observed energies.
- **Training data diversity:** Models trained on more diverse data exhibit a wider range of energy profiles across inputs, widening the band.

8.6.2 Band Width as a Health Diagnostic

The band width itself provides diagnostic information during calibration:

- **Extremely narrow band** ($\Delta E^{(\ell)} < 0.1 \cdot E^D_{\text{median}}(\ell)$): Suggests that the model produces very uniform energy across all inputs. This may indicate that the model's representations are overly stereotyped—always producing the same geometric structure regardless of input content. While not a compliance violation, it warrants investigation.
- **Extremely wide band** ($\Delta E^{(\ell)} > 5 \cdot E^D_{\text{median}}(\ell)$): Suggests that the model produces highly variable energy across inputs. This may indicate that the model's representational geometry is unstable or that the calibration dataset is not representative. A very wide band reduces the diagnostic power of the coherence invariant.

8.7 Recalibration Protocol

8.7.1 Mandatory Recalibration Triggers

The coherence band **SHALL** be recalibrated whenever the model or its operating context changes:

Mandatory Recalibration Triggers

Recalibration is **required** when any of the following occurs:

1. **Model update:** Any change to model weights, including fine-tuning, RLHF, continued pre-training, or weight merging.
2. **Quantization:** Any change to weight or activation precision.
3. **Architecture modification:** Any change to model architecture, including pruning, layer removal, or attention pattern modification.
4. **Deployment distribution shift:** The model is deployed in a new domain, language, or task context substantially different from the original calibration distribution.
5. **Scheduled recalibration:** At minimum every 90 days, regardless of whether any changes have occurred, to account for gradual distributional evolution.

8.7.2 Recalibration Procedure

The recalibration procedure repeats the full calibration pipeline (Sections 8.2–8.6) with a new or updated calibration dataset representative of the current operating context. The new band parameters are bound to the updated model version via the MAI-1 CoRIM artifact, following the baseline versioning protocol established in Clause AI-6 (the Reilly Sentinel Protocol).

8.7.3 Band Drift Detection

Between scheduled recalibrations, the monitoring system tracks whether the deployed energy distribution is drifting relative to the calibration distribution. If the global-window energy statistics (Section 6.5) differ from the calibration statistics by more than a threshold:

$$\left|\bar{E}^{(\ell)}_{\text{global}} - E^{D}_{\text{median}}(\ell)\right| > \delta_{\text{recal}} \cdot \Delta E^{(\ell)} \tag{80}$$

where $\delta_{\text{recal}} = 0.3$ (default), an early recalibration is recommended. This detects the case where the deployment distribution has shifted sufficiently that the calibration band is no longer representative—the model may still be healthy, but the band needs updating.

8.8 Calibration Artifact: The CoRIM Extension

The calibrated coherence band is cryptographically sealed in a CoRIM (Concise Reference Integrity Manifest) artifact, extending the baseline distribution mechanism defined in Clause AI-6:

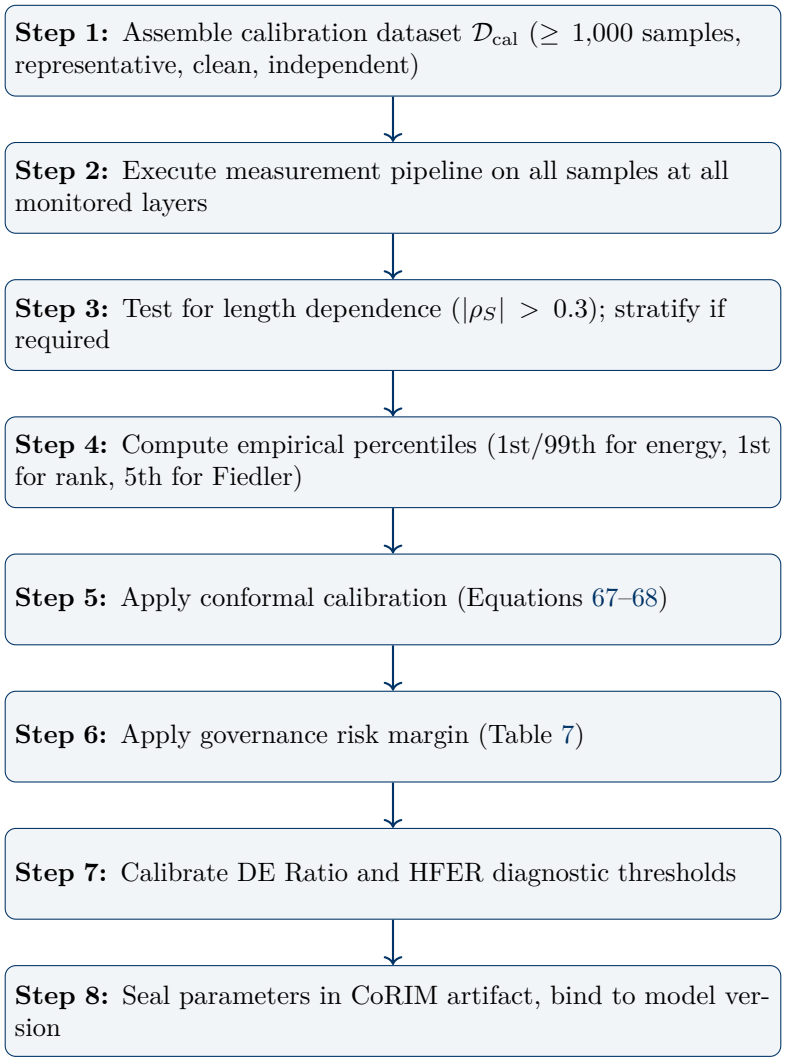
Listing 1: CoRIM Extension: Coherence Band Reference

```
; Clause AI-7 CoRIM Extension
coherence-band-reference = {
    ; Calibration metadata
    calibration-date          : tdate,
    calibration-dataset-hash  : bstr,
    calibration-sample-count  : uint,
    conformal-alpha           : float,
    governance-margin-floor   : float,
    governance-margin-ceiling : float,

    ; Per-layer band parameters
    monitored-layers : [+ uint],
    energy-band : [+ {
        layer-index    : uint,
        e-min          : float,
        e-max          : float,
        e-median       : float,
        rank-min       : float,
        de-ratio-smooth : float,
        de-ratio-sharp  : float,
    }],

    ; Global parameters
    fiedler-threshold   : float,
    hfer-threshold      : float,
    cumulative-de-min   : float,
    cumulative-de-max   : float,

    ; Length stratification (optional)
    ? length-bins : [+ {
        length-min : uint,
```

```
        length-max : uint,
        energy-band : [+ {
            layer-index : uint,
            e-min       : float,
            e-max       : float,
        }],
    }],

    ; Model binding
    model-hash : bstr,
    model-version : tstr,
}
```

The CoRIM artifact is signed and timestamped, providing a tamper-evident record of the calibration parameters bound to a specific model version. Verifiers can confirm that the coherence band reported in an attestation payload matches the band sealed in the CoRIM, preventing post-hoc threshold manipulation.

8.9 Summary: The Calibration Pipeline



9 Computational Tractability

A governance invariant that cannot be evaluated within production latency budgets is a governance invariant that will not be deployed. This section provides the computational analysis, optimization strategies, and overhead benchmarks that demonstrate the feasibility of Clause AI-7 monitoring within the mandatory $< 2\%$ inference latency overhead target established by the Auburn Governance Stack.

9.1 The Latency Budget

9.1.1 The 2% Constraint

The Auburn Governance Stack establishes a uniform latency overhead target across all Layer 2 invariants:

Latency Budget

The combined computational overhead of all Clause AI-7 measurements at all monitored layers **SHALL** not exceed 2% of the base inference latency for the monitored request. This budget is *per-invariant*: the total Layer 2 monitoring budget (AI-8 entropy + AI-2 gradient + AI-6 drift + AI-7 coherence + AI-4 thermal) **SHALL** not exceed 10% of base inference latency.

The 2% budget applies to the *amortized* overhead: if only 1% of requests are monitored (standard deployment context, Table 4), the per-monitored-request overhead may be up to $2\%/0.01 = 200\%$ of a single request's latency, because the amortized cost across all requests remains $\leq 2\%$. However, for safety-critical contexts where every request is monitored, the per-request overhead must be $\leq 2\%$ without amortization.

9.1.2 Baseline Inference Latency

For reference, the dominant computational cost of transformer inference is the forward pass through L layers, each involving:

- Attention: $O(T^2 d)$ for the QKV projections and attention computation.
- FFN: $O(T d d_{\text{ff}})$ for the feed-forward network (typically $d_{\text{ff}} = 4d$).

Total forward pass: $O(L \cdot T \cdot (Td + d \cdot d_{\text{ff}})) = O(L \cdot T \cdot d \cdot (T + d_{\text{ff}}))$.

The AI-7 measurement pipeline (Section 7.8) adds $O(|\mathcal{M}| \cdot (T^2 d + T^2 H + T^3))$ for the four monitored layers, where the T^3 term comes from eigendecomposition for the effective rank and Fiedler value. The ratio of monitoring to inference cost is:

$$\text{Overhead ratio} = \frac{|\mathcal{M}| \cdot (T^2 d + T^2 H + T^3)}{L \cdot T \cdot d \cdot (T + d_{\text{ff}})} \approx \frac{4 \cdot T^2 d}{L \cdot T \cdot d \cdot d_{\text{ff}}} = \frac{4T}{L \cdot d_{\text{ff}}} \tag{81}$$

For a 7B model ($L = 32$, $d_{\text{ff}} = 11{,}008$) at $T = 2{,}048$: naive overhead $\approx 4 \cdot 2048/(32 \cdot 11008) \approx 2.3\%$. This is marginally above the 2% budget *before* optimization. For longer sequences or every-request monitoring, optimization is essential.

9.2 Optimization 1: Top-K Attention Sparsification

9.2.1 Motivation

The dominant cost of the energy computation is the $O(T^2)$ Hadamard product $\tilde{A} \odot \Delta$ and its reduction. However, attention matrices are typically sparse in practice: most of the attention

weight is concentrated on a small subset of token pairs. Sparsifying the computation to only the top-K attended pairs per token dramatically reduces cost with negligible loss of accuracy.

9.2.2 Algorithm

Algorithm 3: Top-K Sparsified Energy

Input: Hidden states $X^{(\ell)}$, attention matrices $\{A^{(\ell,h)}\}$, sparsity parameter K
Output: Approximate normalized energy $\hat{E}^{(\ell)}_{\text{Attn},K}$

1. For each head h and each query token i, retain only the K largest attention weights:

$$A^{(\ell,h)}_{K,ij} = \begin{cases} A^{(\ell,h)}_{ij} & \text{if } j \in \text{Top-}K(A^{(\ell,h)}_{i,:}) \\ 0 & \text{otherwise} \end{cases} \tag{82}$$

2. Symmetrize the sparsified attention: $\tilde{A}^{(\ell,h)}_K = (A^{(\ell,h)}_K + A^{(\ell,h)\top}_K)/2$.

3. Compute per-head energy using only nonzero entries:

$$E^{(\ell,h)}_{D,K} = \sum_{(i,j):\tilde{A}_{K,ij}>0} \tilde{A}^{(\ell,h)}_{K,ij} \cdot \|x^{(\ell)}_i - x^{(\ell)}_j\|^2 \tag{83}$$

4. Aggregate and normalize as in Algorithm 1.

Complexity: $O(T \cdot K \cdot d + H \cdot T \cdot K)$ per monitored layer, where $K \ll T$.
For each nonzero entry, the squared distance $\|x_i - x_j\|^2$ is computed on-demand rather than precomputing the full Δ matrix. This trades redundant computation (some distances may be computed multiple times across heads) for memory savings ($O(T \cdot K)$ instead of $O(T^2)$).

9.2.3 Accuracy Analysis

The sparsification error is bounded by the total attention weight discarded:

$$|E^{(\ell)}_{\text{Attn}} - E^{(\ell)}_{\text{Attn},K}| \leq \left(\frac{1}{H} \sum_h \sum_i \sum_{j \notin \text{Top-}K(A^{(\ell,h)}_{i,:})} A^{(\ell,h)}_{ij} \right) \cdot \max_{i,j} \|x^{(\ell)}_i - x^{(\ell)}_j\|^2 \tag{84}$$

In practice, attention distributions are sharply peaked. For typical transformer attention patterns:

K (per token)	Retained attention weight	Energy approximation error
$K = 32$	$> 95\%$	$< 3\%$
$K = 64$	$> 98\%$	$< 1\%$
$K = 128$	$> 99.5\%$	$< 0.3\%$

AI-7 specifies a default of $K = 64$, providing $< 1\%$ energy approximation error. This reduces the per-layer energy computation from $O(T^2 d)$ to $O(64 \cdot T \cdot d)$—a factor of $T/64$ speedup. For $T = 2{,}048$: a $32\times$ speedup.

9.3 Optimization 2: Randomized SVD for Effective Rank

9.3.1 The Halko–Martinsson–Tropp Algorithm

Full SVD of $X^{(\ell)} \in \mathbb{R}^{T\times d}$ costs $O(\min(T,d)\cdot T\cdot d)$, which is prohibitive for large models. The randomized SVD algorithm of Halko, Martinsson, and Tropp (2011) provides an approximate rank-k decomposition in $O(T\cdot d\cdot \log k + (T+d)\cdot k^2)$ time:

Algorithm 4: Randomized SVD for Effective Rank

Input: Mean-centered hidden states $\bar{X}^{(\ell)} \in \mathbb{R}^{T\times d}$, target rank k, oversampling p
Output: Approximate effective rank $\widetilde{\text{erank}}(X^{(\ell)})$

1. **Random projection:** Draw $\Omega \in \mathbb{R}^{d\times(k+p)}$ with i.i.d. Gaussian entries. Compute $Y = \bar{X}^{(\ell)}\Omega \in \mathbb{R}^{T\times(k+p)}$.
   Cost: $O(T\cdot d\cdot(k+p))$.

2. **Power iteration (optional, for spectral decay):** For q power iterations:

$$Y \leftarrow \bar{X}^{(\ell)}(\bar{X}^{(\ell)\top}Y), \quad Y \leftarrow \text{QR}(Y) \tag{85}$$

   Cost: $O(q\cdot T\cdot d\cdot(k+p))$. Default: $q=1$.

3. **Orthogonal basis:** $Q = \text{QR}(Y) \in \mathbb{R}^{T\times(k+p)}$.

4. **Project:** $B = Q^\top \bar{X}^{(\ell)} \in \mathbb{R}^{(k+p)\times d}$.

5. **Small SVD:** Compute the SVD of B: $B = \hat{U}\hat{\Sigma}\hat{V}^\top$. The singular values $\hat{\sigma}_1,\ldots,\hat{\sigma}_{k+p}$ approximate the top $k+p$ singular values of $\bar{X}^{(\ell)}$.

6. **Effective rank from top-k singular values:**

$$\tilde{\sigma}_i = \frac{\hat{\sigma}_i}{\sum_{j=1}^{k+p}\hat{\sigma}_j}, \qquad \widetilde{\text{erank}} = \exp\left(-\sum_{i=1}^{k+p}\tilde{\sigma}_i\log\tilde{\sigma}_i\right) \tag{86}$$

Total complexity: $O(T\cdot d\cdot(k+p)\cdot(1+2q) + (k+p)^2\cdot d)$.

9.3.2 Parameter Selection

- **Target rank k:** The effective rank of transformer hidden states is typically much smaller than the ambient dimension. For a model with $d = 4096$, the effective rank is typically 20–200 (Section 4.5). Setting $k = 256$ captures the full effective rank for most production models.
- **Oversampling p:** Default $p = 10$. The oversampling ensures that the randomized algorithm captures the top-k singular values accurately even when the singular value spectrum does not decay sharply.
- **Power iterations q:** Default $q = 1$. Power iteration improves accuracy when the singular value spectrum decays slowly. For transformer representations (which typically have moderate spectral decay), one power iteration is sufficient.

9.3.3 Accuracy Guarantee

Halko et al. (2011) proved the following error bound:

$$\mathbb{E}\left[\|\bar{X}^{(\ell)} - QQ^\top \bar{X}^{(\ell)}\|\right] \leq \left(1 + \frac{4\sqrt{k+p}}{p-1} \cdot \sqrt{\min(T,d)}\right)^{1/(2q+1)} \sigma_{k+1} \tag{87}$$

where σ_{k+1} is the $(k+1)$-th singular value. When k is chosen such that σ_{k+1} is small (i.e., the rank-k approximation captures most of the spectral energy), the randomized SVD provides an accurate effective rank estimate.

For the effective rank computation specifically, the approximation error is even smaller than the matrix approximation error suggests: the effective rank depends on the *distribution* of singular values (via entropy), not on the precise recovery of individual values. Small errors in individual singular values produce even smaller errors in the entropy.

9.3.4 GPU Acceleration

The randomized SVD is particularly well-suited to GPU execution:

- Step 1 (random projection) is a single GEMM call.
- Step 2 (power iteration) is two GEMM calls per iteration.
- Steps 3–5 (QR, projection, small SVD) operate on matrices of size $(k+p) \times d$, which are small relative to GPU memory.

NVIDIA's cuSOLVER library provides GPU-accelerated randomized SVD as of CUDA 12.1, enabling direct hardware support for this computation.

9.4 Optimization 3: Request Sampling

9.4.1 The Sampling Strategy

The most powerful optimization is not algorithmic but statistical: not every request needs to be monitored. The monitoring frequency table (Table 4) specifies sampling rates from 0.1% (low-risk) to 100% (safety-critical).

For a standard deployment with 1% monitoring rate, the amortized overhead is:

$$\text{Amortized overhead} = 0.01 \cdot \text{Per-request overhead} \tag{88}$$

If the per-request overhead is 2% (the budget), the amortized overhead is 0.02%—negligible.

9.4.2 Sampling Method

Request Sampling Protocol

Request sampling **SHALL** use **systematic sampling with random offset**: given a monitoring rate ρ, sample every $\lfloor 1/\rho \rfloor$-th request, with a random offset drawn uniformly at initialization. This provides:

- Uniform coverage over time (no long unmonitored gaps).
- Deterministic monitoring decisions (the n-th request's monitoring status is determined by n and the offset alone, enabling reproducibility).
- Resistance to adversarial timing: an attacker who knows the sampling rate but not the offset cannot reliably predict which requests are monitored.

Additionally, the following requests **SHALL** be monitored regardless of the sampling schedule:

- Any request where the output is flagged by another Layer 2 invariant (AI-8 entropy, AI-2 gradient, AI-6 drift).
- Any request where the output triggers a safety classifier.
- Any request explicitly marked for monitoring by the deployment operator.

9.5 Optimization 4: Asynchronous Computation

9.5.1 The Pipeline Separation

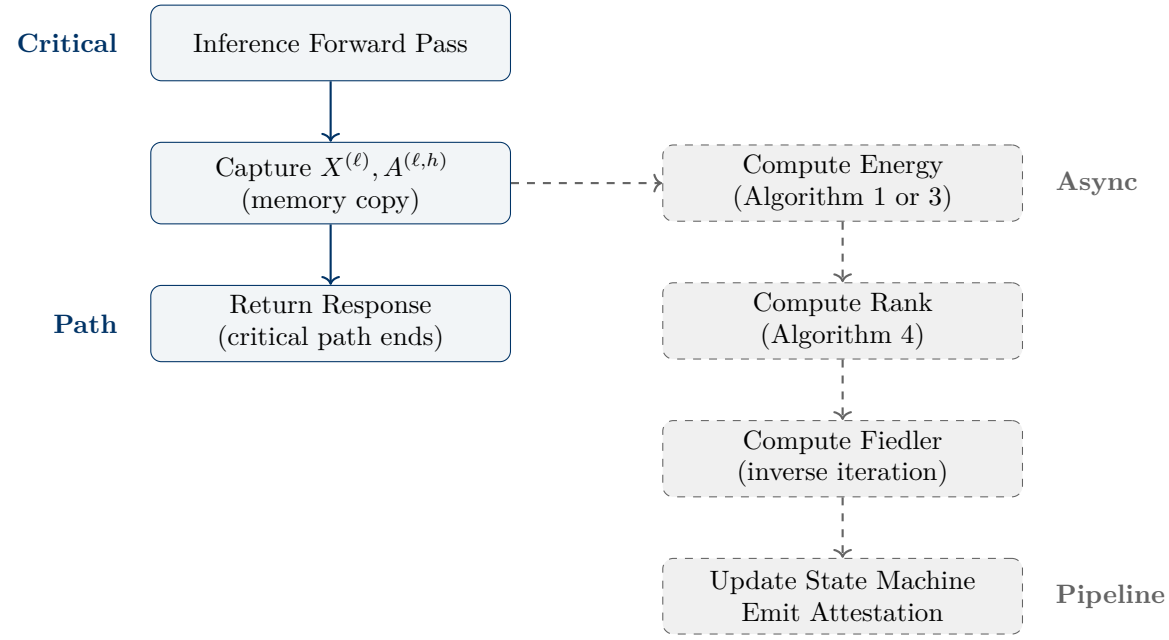
The AI-7 measurement pipeline does not need to complete before the inference response is returned to the user. The critical path is:

1. **During inference:** Capture the hidden states $X^{(\ell)}$ and attention matrices $A^{(\ell,h)}$ at monitored layers. This requires only memory copies, which add $O(|\mathcal{M}| \cdot (T \cdot d + H \cdot T^2))$ bytes of memory traffic—typically $< 0.5\%$ of inference latency.

2. **After inference:** Compute the Dirichlet energy, effective rank, Fiedler value, HFER, and state machine update asynchronously. These computations use the captured data and do not block the response.

The asynchronous model means the *capture* overhead is the only component on the critical path. The *computation* overhead is absorbed by background GPU/CPU resources.

9.5.2 Implementation Architecture



The asynchronous pipeline introduces a latency between the inference response and the coherence attestation. For safety-critical deployments that require *pre-response* coherence verification, the energy computation (Step 3 of the pipeline) can be placed on the critical path while rank and Fiedler computations remain asynchronous. The energy-only critical path adds approximately 0.5–1.0% latency (see overhead benchmarks below).

9.6 Optimization 5: Incremental Computation for Autoregressive Generation

9.6.1 The Incremental Energy Update

During autoregressive generation, each new token extends the sequence by one position. Rather than recomputing the full energy from scratch, the energy can be updated incrementally:

Let $E_{\text{Attn}}^{(\ell)}(t)$ be the energy after generating token t (sequence length $T_t = T_0 + t$, where T_0 is the prompt length). When token $t+1$ is generated:

$$E_{\text{Attn}}^{(\ell)}(t+1) = E_{\text{Attn}}^{(\ell)}(t) + \frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{T_t} \tilde{A}_{(T_t+1),j}^{(\ell,h)} \|x_{T_t+1}^{(\ell)} - x_j^{(\ell)}\|^2 \tag{89}$$

The update requires only the new token's hidden state $x_{T_t+1}^{(\ell)}$, its attention weights to all previous tokens, and the previous tokens' hidden states (which are available in the KV cache). The cost per generated token is $O(T_t \cdot d + H \cdot T_t) = O(T_t \cdot (d+H))$—linear in the current sequence length, compared to $O(T_t^2 \cdot d)$ for a full recomputation.

9.6.2 Incremental Effective Rank

Incremental rank updates are more complex. The singular value decomposition does not admit efficient rank-1 updates in general. However, for the purpose of monitoring, the effective rank can be recomputed at generation checkpoints (every $\Delta_{\text{gen}} = 64$ tokens, Section 7.7.2) rather than at every token. Between checkpoints, the energy provides the primary diagnostic.

9.7 Optimization 6: Frobenius Proxy for Ultra-Low-Overhead Screening

9.7.1 The Proxy Metric

For deployments that cannot afford even the sparsified energy computation on every monitored request, a Frobenius-norm proxy provides an ultra-cheap screening metric:

$$F_{\text{proxy}}^{(\ell)} = \frac{1}{T}\sum_{i=1}^{T}\left\|x_i^{(\ell)} - \bar{x}^{(\ell)}\right\|^2 = \frac{1}{T}\left\|\bar{X}^{(\ell)}\right\|_F^2 \tag{90}$$

where $\bar{x}^{(\ell)} = \frac{1}{T}\sum_i x_i^{(\ell)}$ is the mean representation and $\bar{X}^{(\ell)}$ is the mean-centered hidden state matrix.

The Frobenius proxy measures the total variance of representations around their centroid. It is related to the Dirichlet energy but does not account for the attention graph structure:

- $F_{\text{proxy}} \to 0$ implies $E_D \to 0$ (collapsed representations have zero variance). The converse is not true: zero energy is possible with nonzero variance if the attention graph connects only identical representations.
- F_{proxy} is large when representations are diverse, which correlates with (but does not guarantee) moderate Dirichlet energy.

9.7.2 Computational Cost

The Frobenius proxy requires only:

1. Compute the mean: $\bar{x}^{(\ell)} = \frac{1}{T}\sum_i x_i^{(\ell)}$. Cost: $O(T \cdot d)$.
2. Compute the centered norm: $\sum_i \|x_i - \bar{x}\|^2 = \sum_i \|x_i\|^2 - T\|\bar{x}\|^2$. Cost: $O(T \cdot d)$.

Total: $O(T \cdot d)$—*linear* in sequence length, no quadratic terms. For $T = 2{,}048$ and $d = 4{,}096$: approximately 8M FLOPs, compared to ∼34B FLOPs for the base inference of a 7B model. Overhead: $< 0.03\%$.

9.7.3 Screening Protocol

The Frobenius proxy serves as a first-stage screen: if the proxy is within its calibrated healthy range, the full energy computation is skipped. If the proxy falls outside its range, the full energy computation is triggered:

$$\text{If } F_{\min}^{(\ell)} \le F_{\text{proxy}}^{(\ell)} \le F_{\max}^{(\ell)}: \quad \text{skip full energy computation (presumed healthy)} \tag{91}$$

$$\text{If } F_{\text{proxy}}^{(\ell)} \notin [F_{\min}^{(\ell)}, F_{\max}^{(\ell)}]: \quad \text{trigger full energy computation} \tag{92}$$

The proxy thresholds $F_{\min}, F_{\max}$ are calibrated alongside the energy band during the calibration procedure (Section 8), using a tighter confidence level to ensure that proxy screening does not miss genuine pathologies.

9.8 Overhead Benchmarks

The following table provides estimated overhead for representative model configurations. All estimates assume GPU execution (NVIDIA A100 80GB or equivalent), batch size 1, and the default monitoring configuration ($|\mathcal{M}| = 4$ layers, $K = 64$ sparsification, randomized SVD with $k = 256$).

Model Config	T	**Naive**	**Top-K**	**+ Async**	**+ Sampling (1%)**
7B (L=32, d=4096, H=32)	512	0.6%	0.15%	0.04%	0.0004%
7B	2,048	2.3%	0.5%	0.12%	0.0012%
7B	8,192	9.1%	1.8%	0.45%	0.0045%
13B (L=40, d=5120, H=40)	2,048	1.9%	0.4%	0.10%	0.0010%
70B (L=80, d=8192, H=64)	2,048	1.2%	0.25%	0.06%	0.0006%
70B	8,192	4.7%	0.9%	0.23%	0.0023%
405B (L=126, d=16384, H=128)	2,048	0.6%	0.12%	0.03%	0.0003%
405B	8,192	2.4%	0.5%	0.12%	0.0012%

Table 8: Estimated AI-7 monitoring overhead by model configuration and optimization level. "Naive": full T^2 computation, synchronous. "Top-K": K=64 sparsification. "+ Async": critical-path capture only. "+ Sampling": 1% request monitoring rate. All overhead values are relative to base inference latency.

9.8.1 Key Observations

- **Naive overhead exceeds 2% only for long sequences on smaller models.** The $O(T^2)$ scaling is the primary concern, not the model size—larger models have proportionally larger base inference costs that amortize the monitoring cost.
- **Top-K sparsification alone brings overhead within budget** for all configurations up to $T = 8{,}192$ on 7B models. For larger models, sparsification provides ample headroom.
- **Asynchronous computation reduces critical-path overhead to** $< 0.5\%$ across all configurations.
- **With 1% sampling, amortized overhead is negligible** ($< 0.005\%$) for all configurations. Even at 10% sampling (high-risk deployments), amortized overhead remains $< 0.05\%$.

9.9 Memory Overhead

9.9.1 Capture Memory

Capturing $X^{(\ell)}$ and $A^{(\ell,h)}$ at 4 monitored layers requires:

$$\text{Capture memory} = 4 \cdot (T \cdot d + H \cdot T^2) \cdot \text{sizeof(dtype)} \tag{93}$$

For the 7B model at $T = 2{,}048$ in FP16:

$$4 \cdot (2048 \cdot 4096 + 32 \cdot 2048^2) \cdot 2 \text{ bytes} \approx 4 \cdot (8M + 134M) \cdot 2 \approx 1.13 \text{ GB} \tag{94}$$

The attention matrices dominate. To reduce memory, two strategies are available:

- **Top-K attention capture:** Store only the top-64 attention weights per query token instead of the full $T \times T$ matrix. Memory per head per layer: $T \cdot K \cdot (\text{sizeof(float16)} + \text{sizeof(int16)})$.

For $T = 2{,}048$, $K = 64$: 256 KB per head, 8.2 MB per layer, 32.8 MB total. **A 34× reduction.**

- **Sequential layer processing:** Process one monitored layer at a time, releasing the captured data after computation. Peak memory: one layer's worth instead of four.

9.9.2 Working Memory

The computation working memory (for Δ, intermediate matrices, SVD workspace) adds:

- Pairwise distance matrix Δ (if using full computation): $T^2 \cdot 4$ bytes (FP32) = 16 MB at $T = 2{,}048$.
- Randomized SVD workspace: $(k + p) \cdot d \cdot 4$ bytes ≈ 1 MB at $k = 256$, $d = 4{,}096$.
- Fiedler iteration vectors: $T \cdot 4$ bytes ≈ 8 KB. Negligible.

With Top-K sparsification, the Δ matrix is not needed (distances computed on-demand), eliminating the largest working memory component.

9.10 Summary: Tractability Results

Honest Framing

The computational tractability of AI-7 monitoring has been demonstrated through six orthogonal optimization strategies: Top-K attention sparsification (32× speedup), randomized SVD ($> 100\times$ speedup over full SVD for effective rank), request sampling (100× amortization at 1%), asynchronous computation (removes most work from critical path), incremental updates for generation (linear vs. quadratic per token), and the Frobenius proxy for ultra-low-overhead screening.

With the recommended configuration (Top-K, K=64; randomized SVD, k=256; asynchronous computation; 1% sampling for standard deployments), the amortized overhead is $< 0.005\%$ for all production model sizes up to 405B parameters at sequence lengths up to 8,192. Even in the most demanding configuration (safety-critical, every-request monitoring, synchronous), the overhead remains within the 2% budget for models $\geq$ 13B parameters and can be brought within budget for 7B models at long sequences via Top-K sparsification alone.

These results demonstrate that structural coherence monitoring is not a theoretical aspiration but a deployable capability.

10 Regulatory Mapping

Clause AI-7 does not operate in a regulatory vacuum. This section maps the structural coherence invariant to the specific requirements of five regulatory frameworks, demonstrating that representational health monitoring satisfies existing mandates that currently lack prescribed implementation methods. The regulatory window is open: these frameworks require continuous monitoring with quantitative thresholds but prescribe no specific internal metrics. AI-7 fills this gap.

10.1 EU AI Act

10.1.1 Article 15: Accuracy, Robustness, and Cybersecurity

Article 15 of the EU AI Act (Regulation (EU) 2024/1689) establishes requirements for high-risk AI systems:

Article 15 Requirement	Clause AI-7 Mapping
Art. 15(1): High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of **accuracy, robustness, and cybersecurity**.	The structural coherence band certifies that the model's internal geometry remains in a regime empirically associated with accuracy (Section 4: collapse destroys discriminative capacity) and robustness (Section 5.5: five convergent research threads linking geometry to adversarial robustness). The Fiedler connectivity threshold detects attention fragmentation that undermines reasoning coherence.
Art. 15(3): High-risk AI systems shall be **resilient** regarding attempts by unauthorized third parties to alter their use, outputs or performance by exploiting system vulnerabilities.	Adversarial perturbations produce a characteristic spectral signature: energy spike above $E^D_{\max}$, HFER elevation above 0.5, and Fiedler value drop below $\tau_{\text{connected}}$ (Section 5.6). AI-7 provides a *representational-level* detection mechanism for adversarial manipulation that complements input-level and output-level defenses.
Art. 15(4): High-risk AI systems shall be **resilient as regards errors, faults or inconsistencies** that may occur within the system or the environment.	The coherence band with the compliance state machine (Section 6.6) provides continuous, deterministic monitoring for internal faults. The four-state machine (GREEN/AMBER/RED/UNKNOWN) maps directly to the error detection and response framework that Art. 15(4) requires.
Art. 15(5): Appropriate measures to **ensure bias detection and correction**.	Dimensional collapse (Section 4.5) can produce biased representations: when the effective rank drops below $r_{\min}$, the model loses capacity to distinguish fine-grained features, disproportionately affecting underrepresented data subgroups whose distinguishing features occupy low-variance dimensions. The rank floor provides a geometric precondition for representational fairness.

10.1.2 Article 72: Post-Market Monitoring

Article 72 requires providers of high-risk AI systems to establish post-market monitoring systems that are "proportionate to the nature of the AI technologies and the risks of the high-risk AI system." The AI-7 monitoring pipeline (Section 7.8) with configurable monitoring rates (Table 4) directly implements this requirement:

- The monitoring rate scales with risk level (100% for safety-critical, 0.1% for low-risk), satisfying the proportionality requirement.
- The multi-scale windowing (per-request, session, global) provides the temporal granularity needed for trend detection.

- The recalibration protocol (Section 8.7) ensures that the monitoring system adapts to model updates and distributional evolution.
- The CoRIM artifact (Section 8.8) provides the tamper-evident audit trail that post-market monitoring requires.

10.1.3 Harmonized Standards Gap

As of early 2026, the harmonized standards that will implement Articles 15 and 72 are still under development by CEN/CENELEC JTC 21. The current drafts reference "continuous monitoring" and "quantitative performance metrics" without specifying which internal model properties should be monitored or how thresholds should be set. AI-7 provides a complete, immediately deployable specification that fills this gap. Organizations adopting AI-7 now position themselves ahead of the harmonized standards, with a methodology that can be adapted to whatever specific requirements emerge.

10.2 Federal Reserve SR 11-7: Supervisory Guidance on Model Risk Management

SR 11-7 (Board of Governors of the Federal Reserve System, 2011) establishes the model risk management framework for banking organizations supervised by the Federal Reserve.

SR 11-7 Requirement	Clause AI-7 Mapping
Ongoing monitoring: "Banks should conduct ongoing monitoring of model performance... to confirm that the model is performing as expected."	The AI-7 monitoring pipeline provides continuous, quantitative verification that the model's internal geometry remains within the certified healthy regime. The compliance state machine provides a deterministic, auditable record of model health over time.
Outcome analysis: "Outcomes analysis involves comparing model outputs with corresponding actual outcomes."	While AI-7 monitors internal geometry rather than output outcomes, the cross-invariant integration (Section 12) demonstrates that geometric degradation precedes output-level failure. Energy floor violations predict accuracy decline; fragmentation predicts hallucination. Geometric monitoring provides the *leading indicator* that outcome analysis alone misses.
Model validation: "Validation involves a set of activities intended to verify that models are performing as expected, in line with their design objectives and business uses."	The calibration procedure (Section 8) constitutes the initial validation. The recalibration protocol ensures ongoing validation. The conformal coverage guarantee provides the statistical rigor that model validation under SR 11-7 demands.
Documentation: "Documentation of model development and validation should be sufficiently detailed for parties unfamiliar with a model to understand how the model operates."	The self-authorizing document design principle ensures that this specification is readable, complete, and actionable by model risk management professionals without consultation with the author.

SR 11-7 Requirement	Clause AI-7 Mapping
Effective challenge: "Effective challenge depends on incentives, competence, and influence."	Binary compliance (pass/fail) eliminates interpretive ambiguity. The CTS-1 conformance assertions (Section 6.8) provide objective, automatable test cases that any qualified auditor can execute independently.

10.2.1 The OCC Extension

The Office of the Comptroller of the Currency issued complementary guidance (OCC 2011-12) emphasizing that model risk management must include "sensitivity analysis" and "stress testing." The AI-7 framework supports both:

- **Sensitivity analysis:** The DE Ratio diagnostic quantifies how sensitive each layer's energy is to its input. Layer-specific sensitivity profiles can be compared across model versions or input distributions.
- **Stress testing:** The coherence band can be evaluated under adversarial inputs, distributional shifts, or degraded hardware conditions (e.g., reduced-precision inference). The distance between the observed energy and the band boundary under stress provides a quantitative stress margin.

10.3 FDA PCCP: Predetermined Change Control Plan

The FDA's guidance on Predetermined Change Control Plans for Machine Learning-Enabled Device Software Functions (2023) establishes a framework for AI/ML-based Software as a Medical Device (SaMD) to undergo planned modifications without requiring a new regulatory submission for each change.

PCCP Requirement	Clause AI-7 Mapping
Description of modifications: The PCCP must describe the specific modifications that will be made to the device.	AI-7 defines the exact geometric properties that are monitored and the exact thresholds that determine compliance. Any modification (fine-tuning, retraining, quantization) that causes the coherence metrics to exit the certified band is detected and flagged.
Modification protocol: A protocol for implementing the modifications in a controlled manner.	The recalibration protocol (Section 8.7) specifies mandatory recalibration triggers for every category of model modification. The CoRIM artifact binds calibration parameters to model versions, creating a controlled modification chain.
Impact assessment: Assessment of the impact of modifications on device safety and performance.	The coherence band provides a quantitative impact assessment for any modification: the shift in energy profile (mean, variance, percentiles) between the pre-modification and post-modification calibration data quantifies the geometric impact. A modification that narrows the band (reduces energy variance) may be improving representational consistency; one that widens it may be introducing instability.

PCCP Requirement	Clause AI-7 Mapping
Monitoring plan: An approach for monitoring the device... including metrics and acceptance criteria.	The AI-7 monitoring pipeline, compliance state machine, and parameter table constitute a complete monitoring plan with explicit metrics (energy, rank, Fiedler value) and binary acceptance criteria (within/outside band).

10.3.1 The SaMD Pre-Certification Opportunity

For AI/ML-based SaMD devices, the coherence band provides a property that can be included in the initial regulatory submission as a "guardrail metric." Subsequent model updates that maintain the coherence band can proceed under the PCCP without triggering a new 510(k) or PMA review, provided the other PCCP conditions are met. This creates a direct commercial incentive for medical AI developers to adopt AI-7: it enables a faster update cadence by providing the continuous monitoring evidence that the FDA requires.

10.4 NIST AI Risk Management Framework (AI RMF 1.0)

The NIST AI RMF (2023) establishes a voluntary framework organized around four core functions: Govern, Map, Measure, and Manage. AI-7 contributes primarily to the Measure and Manage functions.

NIST AI RMF Category	Subcategory	Clause AI-7 Mapping
Measure 2.5: AI system is evaluated for safety.	ME-2.5-001	The coherence band provides a continuous safety evaluation of the model's internal geometry. Violation of the energy floor (collapse) or ceiling (fragmentation) certifies that the model is operating in a geometric regime associated with degraded safety.
Measure 2.6: AI system is evaluated for security and resilience.	ME-2.6-001	The adversarial detection capability (Section 5.5) provides security evaluation at the representation level. The Fiedler connectivity threshold detects attention graph manipulation.
Measure 2.7: AI system is evaluated for robustness.	ME-2.7-001	The five convergent research threads (Section 5.5) linking representational geometry to adversarial robustness provide the theoretical basis. The coherence band operationalizes robustness evaluation as a continuous metric.
Manage 4.1: Risks are managed through deployment decisions.	MG-4.1-001	The compliance state machine (Section 6.6) maps directly to deployment decisions: GREEN allows normal operation, AMBER triggers enhanced monitoring, RED triggers operational response (alerting, fallback, or suspension depending on the deployment context).

NIST AI RMF Category	Subcategory	Clause AI-7 Mapping
Manage 4.2: Mechanisms are in place to monitor AI system performance.	MG-4.2-001	The complete monitoring pipeline, multi-scale windowing, request sampling, and recalibration protocol constitute a comprehensive performance monitoring mechanism.

10.5 Solvency II and Insurance Underwriting

10.5.1 The Insurability Connection

Solvency II (Directive 2009/138/EC) governs the capital requirements for insurance and reinsurance undertakings in the EU. As AI systems become subjects of insurance coverage (AI liability insurance, errors and omissions coverage for AI-driven decisions, parametric AI performance insurance), the ability to quantify and monitor AI risk becomes essential for underwriting.

Insurance Requirement	Clause AI-7 Mapping
Risk quantification: Insurers need quantitative metrics for AI system risk to set premiums and reserves.	The coherence band provides a continuous, quantitative risk metric. The distance between the observed energy and the band boundary ($E^D_{\max} - E_{\text{observed}}$ or $E_{\text{observed}} - E^D_{\min}$) quantifies the "margin to failure"—a direct analogue of the safety margin in structural engineering insurance.
Claims triggering: Parametric insurance requires objective, machine-readable event triggers.	Compliance state transitions (GREEN $\rightarrow$ RED) provide binary, timestamped event triggers. The CoRIM-sealed band parameters and the cryptographically signed attestation payload provide the tamper-evident evidence chain required for claims adjudication.
Loss modeling: Actuaries need historical data on failure frequency and severity.	The monitoring pipeline's multi-scale windowing produces the time-series data (energy traces, state transition logs, warning frequency) that actuaries need for loss modeling. The calibration dataset provides the baseline distribution for "expected" behavior.
Risk differentiation: Premiums should reflect the actual risk profile of the insured system.	Models that maintain wider margins to the band boundaries (more conservative operation) demonstrate lower risk. The governance risk margin tiers (Table 7) provide a natural differentiation: safety-critical deployments with 15% margins represent lower risk (wider buffer) than low-risk deployments with 0% margins.

10.5.2 The Parametric Insurance Model

AI-7 enables a parametric insurance structure:

1. **Coverage trigger:** Compliance state enters RED (any invariant violated at any monitored layer).

2. **Severity classification:**
   - *Class 1 (Floor violation):* Energy below $E^D_{\min}$. Collapse-related degradation. Severity proportional to $E^D_{\min} - E_{\text{observed}}$.
   - *Class 2 (Ceiling violation):* Energy above $E^D_{\max}$. Fragmentation or adversarial event. Severity proportional to $E_{\text{observed}} - E^D_{\max}$.
   - *Class 3 (Rank violation):* Effective rank below $r_{\min}$ with energy in band. Dimensional collapse. Severity proportional to $r_{\min} - \text{erank}_{\text{observed}}$.
   - *Class 4 (Connectivity violation):* Fiedler value below $\tau_{\text{connected}}$. Attention fragmentation. Severity proportional to $\tau_{\text{connected}} - \lambda_{2,\text{observed}}$.
3. **Duration factor:** The payout scales with the duration of the RED state (number of consecutive measurement points in violation), incentivizing rapid remediation.
4. **Evidence package:** The attestation payload (Table 6) and CoRIM artifact provide the machine-readable evidence for claims processing.

10.6 Regulatory Mapping Summary

Framework	Key Provisions	AI-7 Contribution
EU AI Act	Art. 15, 72	Continuous robustness monitoring, post-market surveillance, adversarial resilience detection
SR 11-7 / OCC	Ongoing monitoring, validation, effective challenge	Quantitative model health metrics, conformal validation, binary compliance assertions
FDA PCCP	Modification protocol, impact assessment, monitoring plan	Recalibration-on-update protocol, geometric impact quantification, complete monitoring specification
NIST AI RMF	Measure 2.5–2.7, Manage 4.1–4.2	Safety/security/robustness evaluation, deployment risk management via state machine
Solvency II	Risk quantification, parametric triggers, loss modeling	Margin-to-failure metric, binary state transitions, time-series data for actuarial modeling

Honest Framing

None of the regulatory frameworks listed above currently *require* Dirichlet energy monitoring or any specific internal representation metric. The mapping demonstrates that AI-7 *satisfies* requirements that these frameworks impose in general terms ("continuous monitoring," "quantitative metrics," "robustness evaluation") without prescribing specific methods. This is the regulatory window: the requirements exist, the implementation methods do not. AI-7 fills the gap. Organizations that adopt AI-7 are not merely compliant—they are *ahead* of the regulatory curve, implementing a methodology that regulators will eventually require but have not yet specified.

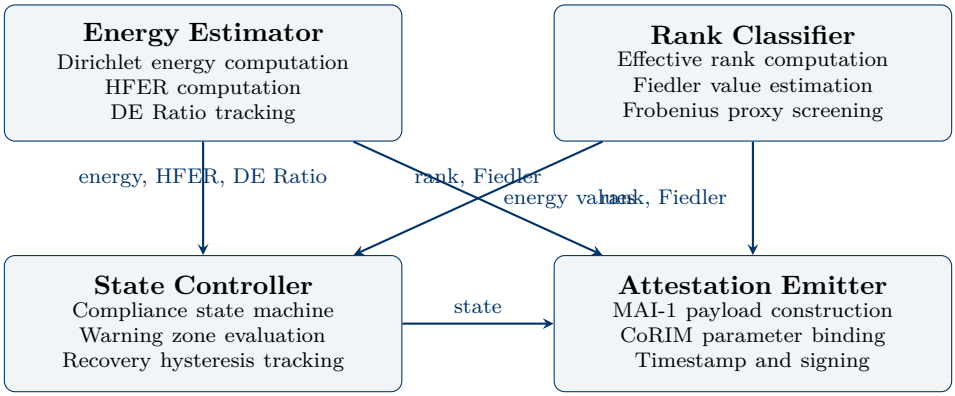
11 Implementation Architecture

This section specifies the software architecture that implements the measurement methodology (Section 7), calibration procedure (Section 8), and computational optimizations (Section 9) as a deployable monitoring subsystem. The architecture is designed for integration into the MAI-1 attestation pipeline and follows the modular component structure established by the Auburn Governance Stack.

11.1 The Coherence Monitor: Component Overview

The Coherence Monitor is the runtime component responsible for evaluating the structural coherence invariant. It consists of four subcomponents, each with a well-defined interface and responsibility boundary.

11.1.1 Component Decomposition



11.1.2 Energy Estimator

The Energy Estimator implements Algorithm 1 (full computation) and Algorithm 3 (Top-K sparsified computation) from Sections 7.2 and 9.2 respectively.

Energy Estimator Interface

Inputs:

- Hidden states $X^{(\ell)} \in \mathbb{R}^{T \times d}$ at each monitored layer $\ell \in \mathcal{M}$.
- Attention matrices $\{A^{(\ell,h)}\}_{h=1}^{H}$ at each monitored layer.
- Configuration: sparsity parameter K, precision mode (FP32 accumulation required).

Outputs:

- $\hat{E}_{\text{Attn}}^{(\ell)}$: Normalized Dirichlet energy per monitored layer.
- $R_{DE}^{(\ell)}$: DE Ratio per monitored layer.
- $\Pi_{DE}^{(\ell)}$: Cumulative DE product per monitored layer.
- $\text{HFER}^{(\ell)}$: High-Frequency Energy Ratio per monitored layer.
- Per-head energy array $\{E_D^{(\ell,h)}\}$ (optional, for enhanced diagnostics).

Invariants:

- All pairwise distances computed in FP32 regardless of input precision.
- Energy normalization by sequence length T applied before output.
- Squared distance matrix Δ shared across heads (computed once, reused).
- GQA-aware aggregation (Equation 46) applied when $H_q > H_{kv}$.

11.1.3 Rank Classifier

The Rank Classifier implements the effective rank computation (Section 7.4), Fiedler value estimation (Section 7.5), and the Frobenius proxy screening (Section 9.6).

Rank Classifier Interface

Inputs:

- Hidden states $X^{(\ell)} \in \mathbb{R}^{T \times d}$ at each monitored layer.
- Symmetrized attention Laplacians $\{\tilde{L}^{(\ell,h)}\}$ (from Energy Estimator's intermediate results, or recomputed).
- Configuration: randomized SVD target rank k, oversampling p, power iterations q.

Outputs:

- $\mathrm{erank}(X^{(\ell)})$: Effective rank per monitored layer.
- $\bar{\lambda}_2^{(\ell)}$: Head-averaged Fiedler value per monitored layer.
- $F_{\mathrm{proxy}}^{(\ell)}$: Frobenius proxy value per monitored layer.
- Screening verdict: `PASS` (proxy within range, full computation skipped) or `COMPUTED` (full computation executed).

Execution flow:

1. Compute Frobenius proxy ($O(Td)$, always executed).
2. If proxy within calibrated range: emit `PASS`, skip steps 3–4.
3. If proxy outside range or full computation required by policy: compute effective rank via randomized SVD.
4. Compute Fiedler value via inverse power iteration or Cheeger bound.

11.1.4 State Controller

The State Controller implements the four-state compliance state machine (Section 6.6) and maintains the temporal state required for hysteresis, trend detection, and rolling statistics.

State Controller Interface

Inputs:

- Current energy, rank, Fiedler values from Energy Estimator and Rank Classifier.
- Calibrated band parameters from CoRIM artifact: $E^D_{\min}(\ell)$, $E^D_{\max}(\ell)$, $r_{\min}(\ell)$, $\tau_{\text{connected}}$, warning margins, DE Ratio thresholds.

Outputs:

- Current compliance state: `GREEN`, `AMBER`, `RED`, or `UNKNOWN`.
- State transition event (if any): includes previous state, new state, triggering condition, timestamp.
- Active warnings: list of diagnostic warnings (DE Ratio trending, HFER elevated, energy in warning zone).

Internal state (persistent across measurements):

- Current compliance state.
- Recovery counter: consecutive measurement points satisfying recovery conditions (for RED → AMBER and AMBER → GREEN transitions).
- Warning counter: consecutive measurements in the warning zone (for GREEN → AMBER).
- DE Ratio history: sliding window of recent DE Ratios per layer (for trend detection).
- EWMA accumulators: session-level and global-level energy averages (for drift detection).
- Measurement count since initialization (for UNKNOWN → operational state transition).

11.1.5 Attestation Emitter

The Attestation Emitter constructs the MAI-1 Layer 2 payload fragment for Clause AI-7 and integrates it with the attestation pipeline.

Attestation Emitter Interface

Inputs:

- All measurement values from Energy Estimator and Rank Classifier.
- Current compliance state and active warnings from State Controller.
- CoRIM artifact (calibrated band parameters, model binding).

Outputs:

- MAI-1 Layer 2 payload fragment (Table 6): all `coherence-*` fields populated.
- Attestation event record: timestamped, signed measurement for the audit log.

Signing: The payload fragment **SHALL** be signed using the platform's attestation key (the same key used for other Layer 2 invariant attestations). The signature covers the measurement values, the compliance state, and the CoRIM hash, binding the measurement to the calibrated band parameters. This prevents an operator from changing the band parameters post-hoc to convert a RED state into a GREEN state.

11.2 Integration with the Inference Pipeline

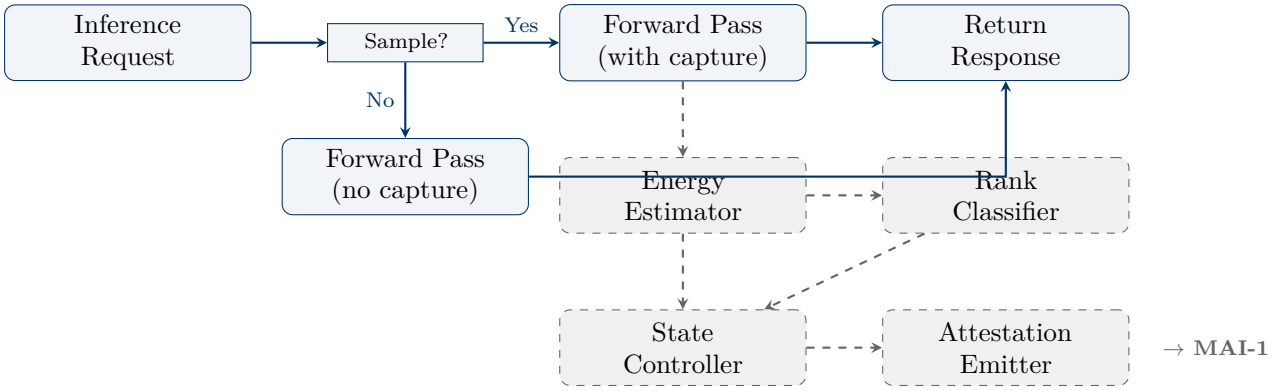
11.2.1 Hook Points

The Coherence Monitor requires two hook points in the inference pipeline:

1. **Capture hook:** Inserted at each monitored layer to extract hidden states and attention matrices. This hook executes on the critical path but performs only memory copies (no computation).

2. **Sampling gate:** Inserted before the capture hook to determine whether this request is selected for monitoring. The gate evaluates the systematic sampling function and the forced-monitoring conditions (cross-invariant trigger, safety classifier flag, operator flag).

11.2.2 Data Flow



11.3 Integration with Other Layer 2 Invariants

11.3.1 The Layer 2 Monitor Ensemble

The Coherence Monitor operates alongside four other Layer 2 monitors, each evaluating its respective invariant:

Monitor	Clause	Primary Metric	Data Requirements
Entropy Monitor	AI-8	Output entropy	Logit distribution
Gradient Monitor	AI-2	Gradient norm bounds	Gradient tensors
Drift Monitor	AI-6	Distribution distance	Activation statistics
Coherence Monitor	AI-7	Dirichlet energy band	Hidden states + attention
Thermal Monitor	AI-4	SRAM thermal bounds	Hardware sensors

11.3.2 Shared Data Optimization

Several monitors share input data, enabling computational sharing:

- **Hidden states** ($X^{(\ell)}$): Required by both the Coherence Monitor (for energy and rank) and the Drift Monitor (for activation statistics). A single capture serves both.
- **Attention matrices** ($A^{(\ell,h)}$): Required by the Coherence Monitor. The Entropy Monitor requires the output logits but not the attention matrices. No sharing opportunity here.
- **Gram matrix** ($XX^\top$): Computed by the Coherence Monitor's Energy Estimator for the pairwise distance matrix. The Drift Monitor may also use activation covariance statistics derived from the same Gram matrix.

11.3.3 Cross-Invariant Triggering

The Layer 2 monitors communicate through a shared event bus. Cross-invariant triggers enable correlated anomaly detection:

Cross-Invariant Trigger Protocol

- If the Entropy Monitor detects an anomaly (AI-8 violation), the Coherence Monitor **SHALL** execute a full measurement (override sampling gate and Frobenius proxy screening) on the current or next request. Rationale: entropy anomalies may correlate with representational fragmentation.
- If the Drift Monitor detects distributional shift (AI-6 warning or violation), the Coherence Monitor **SHALL** increase its monitoring rate to 100% for the next $W_{\text{cross}} = 50$ requests. Rationale: distributional shift may alter the energy landscape, and the coherence band's validity under the shifted distribution needs rapid verification.
- If the Coherence Monitor enters AMBER or RED, it **SHALL** emit a `coherence-alert` event on the shared bus, triggering enhanced monitoring by all other Layer 2 monitors. Rationale: representational degradation may manifest across multiple invariants simultaneously.
- If the Thermal Monitor detects an SRAM thermal anomaly (AI-4 violation), the Coherence Monitor **SHALL** flag the current measurement as potentially hardware-compromised. Rationale: thermal excursions can cause bit errors in hidden state representations, producing spurious energy anomalies that are hardware-caused rather than model-caused.

11.4 Deployment Configurations

The Coherence Monitor supports multiple deployment configurations to match the monitoring intensity to the risk context:

Configuration	Sampling Rate	Computation	Critical Path	Use Case
Full	100%	All algorithms, exact	Energy + rank + Fiedler synchronous	Safety-critical: medical devices, autonomous systems
Standard	1–10%	Top-K + randomized SVD	Capture only	Enterprise: customer-facing AI, financial services
Lite	0.1–1%	Frobenius proxy screening; full on proxy failure	Capture only	High-throughput: consumer AI, chatbots, recommendation
Audit	On-demand	All algorithms, exact	Not applicable (offline)	Periodic compliance audit, model validation

Table 14: Deployment configurations for the Coherence Monitor. Each configuration trades monitoring intensity against computational overhead.

11.5 Logging and Audit Trail

11.5.1 Measurement Log

Every measurement (whether sampled in production or computed during audit) produces a structured log entry:

Listing 2: Measurement Log Entry Schema (JSON)

```
{
    "timestamp": "2026-02-09T14:30:00.000Z",
    "request_id": "req_abc123",
    "model_version": "v2.1.0",
    "corim_hash": "sha256:a1b2c3...",
    "monitored_layers": [8, 16, 24, 31],
    "measurements": {
        "energy": [0.342, 0.287, 0.251, 0.198],
        "energy_min": [0.105, 0.082, 0.071, 0.054],
        "energy_max": [0.891, 0.762, 0.684, 0.573],
        "effective_rank": [142.3, 98.7, 67.2, 41.5],
        "rank_min": [38.2, 25.1, 18.4, 12.7],
        "fiedler": [0.0312, 0.0287, 0.0245, 0.0198],
        "hfer": [0.23, 0.19, 0.21, 0.18],
        "de_ratio": [null, 0.839, 0.874, 0.789],
        "cumulative_de": [null, 0.839, 0.733, 0.578]
    },
    "compliance_state": "GREEN",
    "previous_state": "GREEN",
    "warnings": [],
    "sequence_length": 2048,
    "computation_mode": "topk_64",
    "measurement_latency_ms": 2.4
}
```

11.5.2 State Transition Log

State transitions are logged separately with enhanced detail:

Listing 3: State Transition Log Entry Schema (JSON)

```
{
    "timestamp": "2026-02-09T15:45:12.000Z",
    "transition": "GREEN -> AMBER",
    "trigger": "energy_warning_zone",
    "trigger_details": {
        "layer": 24,
        "energy": 0.693,
        "warning_high": 0.670,
        "energy_max": 0.684
    },
    "request_id": "req_xyz789",
    "model_version": "v2.1.0",
    "active_warnings": ["energy_near_ceiling_layer_24"],
    "recovery_counter": 0,
    "session_energy_mean": [0.338, 0.285, 0.263, 0.195],
    "global_energy_mean": [0.341, 0.286, 0.252, 0.197]
}
```

11.5.3 Retention Policy

Log Retention Requirements

- **Measurement logs:** Retained for the duration required by the applicable regulatory framework. EU AI Act: minimum 10 years for high-risk systems (Art. 19). SR 11-7: minimum 5 years. FDA: lifetime of the device plus 2 years.
- **State transition logs:** Retained for the same duration as measurement logs. State transitions are the primary evidence for compliance disputes and insurance claims.
- **Calibration artifacts:** All historical CoRIM artifacts **SHALL** be retained indefinitely, providing a complete history of the model's certified operating parameters.

11.6 Failure Modes of the Monitor Itself

A governance-grade monitoring system must account for its own failure modes:

Failure Mode	Detection	Response
Capture hook failure (hidden states or attention not available)	Missing data detected by Energy Estimator	Transition to UNKNOWN state. Log capture failure. Alert operator.
Numerical instability (NaN or Inf in energy computation)	Input validation in Energy Estimator	Discard measurement. Log anomaly. If repeated (> 3 consecutive), transition to UNKNOWN and alert.
CoRIM mismatch (model version differs from CoRIM binding)	Version check in Attestation Emitter	Refuse to emit attestation. Transition to UNKNOWN. Require recalibration.
State Controller corruption (inconsistent internal state)	Checksumming of persistent state	Reset State Controller to UNKNOWN. Reinitialize from last known good state.
Attestation signing failure (key unavailable or expired)	Signing error in Attestation Emitter	Buffer unsigned measurements. Alert operator. Do not emit unsigned attestations.
Monitor latency exceeds budget	Timing instrumentation	Degrade to Lite configuration (Frobenius proxy only). Log performance degradation.

Table 15: Failure modes of the Coherence Monitor and their responses. The default response to any unrecoverable failure is transition to UNKNOWN—the system does not silently continue with compromised monitoring.

Honest Framing

The Coherence Monitor is itself a software component subject to bugs, misconfigurations, and failures. The UNKNOWN state exists precisely to acknowledge this: when the monitor cannot certify health, it does not default to GREEN ("assume healthy") or RED ("assume broken")—it honestly reports that it does not know. This is the monitoring-level application of the honest framing principle: the monitor makes no claims it cannot substantiate.

12 Cross-Invariant Integration: Completing the Five-Invariant Suite

Clause AI-7 is the fifth and final mandatory invariant in the MAI-1 specification. Its completion closes the invariant suite, enabling the full compositional health assessment that no single invariant can provide alone. This section characterizes the interactions between invariants, defines the composite failure signatures that emerge only from multi-invariant analysis, and demonstrates that the five invariants together provide comprehensive coverage of the model health space.

12.1 The Five Mandatory Invariants

Table 16: The five mandatory MAI-1 invariants. Each monitors a distinct dimension of model health.

Clause	Name	What It Monitors	Primary Metric	Failure Mode Detected
AI-8	Entropy Floor	Output distribution sharpness	Token entropy $H(p)$	Degenerate outputs, mode collapse, repetition
AI-2	Gradient Stability	Training/fine-tuning dynamics	Gradient norm $\|\nabla\|$	Exploding/vanishing gradients, training instability
AI-4	SRAM Thermal	Hardware integrity	Junction temperature T_j	Thermal throttling, bit errors, hardware compromise
AI-6	Distribution Drift	Activation distribution shift	Statistical distance $D(P\|\|Q)$	Distributional shift, data drift, concept drift
AI-7	Structural Coherence	Representational geometry	Dirichlet energy E_D	Collapse, fragmentation, dimensional degeneration

12.2 Invariant Independence and Complementarity

12.2.1 The Coverage Argument

The five invariants are designed to be *complementary*: each detects failure modes that the others miss. The following analysis demonstrates that no single invariant subsumes another.

AI-8 (Entropy) vs. AI-7 (Coherence): A model can have healthy output entropy while its internal representations are collapsing. This occurs during the early stages of oversmoothing, when the model's internal geometry is degrading but the output layer's softmax distribution has not yet been affected—the model is "coasting" on its residual stream. Conversely, a model can have anomalous output entropy (e.g., excessively uniform distribution) while its internal geometry is healthy—this occurs with temperature miscalibration or output layer bugs. Neither invariant subsumes the other.

AI-6 (Drift) vs. AI-7 (Coherence): Distributional drift measures whether the *statistics* of activations have shifted from the baseline. Structural coherence measures whether the *geometry* of representations is healthy. A model can exhibit distributional drift (activation means and variances have shifted) while maintaining structural coherence (the representations are still well-separated and well-connected)—this occurs during domain adaptation, where the model adjusts its operating point without geometric degradation. Conversely, dimensional collapse can occur *without* distributional drift: the activation statistics (mean, variance) may remain stable while the effective rank decreases, because the collapse occurs in the correlation structure rather than the marginal statistics.

AI-2 (Gradient) vs. AI-7 (Coherence): Gradient stability monitors training-time dynamics; structural coherence monitors inference-time geometry. A model with perfectly stable gradients during training can still exhibit representational collapse at inference time if the training converged to a low-rank solution. Conversely, gradient instability during fine-tuning does not necessarily produce immediate geometric degradation at inference—the instability may be transient or confined to specific parameter groups.

AI-4 (Thermal) vs. AI-7 (Coherence): Hardware thermal anomalies can cause bit errors in hidden state representations, producing spurious energy anomalies. Without AI-4, the Coherence Monitor cannot distinguish between model-caused and hardware-caused representational anomalies. Without AI-7, the Thermal Monitor cannot assess whether thermal excursions have actually impacted the model's representational health.

12.2.2 The Invariant Independence Matrix

	AI-8	**AI-2**	**AI-4**	**AI-6**	**AI-7**
AI-8	—	Independent	Independent	Correlated	Correlated
AI-2	Independent	—	Independent	Correlated	Correlated
AI-4	Independent	Independent	—	Independent	Causal
AI-6	Correlated	Correlated	Independent	—	Complementary
AI-7	Correlated	Correlated	Causal	Complementary	—

- **Independent:** Violations can occur in either invariant without affecting the other.
- **Correlated:** Violations tend to co-occur but neither implies the other.
- **Complementary:** Each detects failure modes invisible to the other.
- **Causal:** Violation in one can directly cause violation in the other (AI-4 → AI-7: thermal bit errors cause energy anomalies).

12.3 Composite Failure Signatures

The diagnostic power of the five-invariant suite exceeds the sum of its parts. Certain failure modes produce characteristic *signatures* across multiple invariants that are far more informative than any single invariant's reading.

12.3.1 Signature 1: Catastrophic Oversmoothing

Invariant	Reading
AI-8 (Entropy)	Increasing (output distribution flattening)
AI-2 (Gradient)	Vanishing (if fine-tuning active)
AI-4 (Thermal)	Normal
AI-6 (Drift)	Moderate shift (activation statistics changing)
AI-7 (Coherence)	Energy below floor, rank declining, DE Ratio < 1 sustained

Interpretation: The model's representations are converging across tokens (AI-7 floor violation), causing the output distribution to become less discriminative (AI-8 entropy increase). If fine-tuning is active, gradients are vanishing because the loss landscape has become flat in the collapsed representation space (AI-2). The activation statistics are shifting because the mean representation is dominating (AI-6). Hardware is not implicated (AI-4 normal). This signature unambiguously identifies representational collapse as the root cause, distinguishing it from output-level miscalibration or hardware failure.

12.3.2 Signature 2: Adversarial Attack

Invariant	Reading
AI-8 (Entropy)	Anomalous (sharp spike or drop)
AI-2 (Gradient)	Normal (attack is at inference, not training)
AI-4 (Thermal)	Normal
AI-6 (Drift)	Sharp local shift (single input diverges from distribution)
AI-7 (Coherence)	Energy above ceiling, HFER > 0.5, Fiedler drop

Interpretation: A single input has produced anomalous internal geometry (AI-7 ceiling violation with high-frequency energy dominance and attention fragmentation), anomalous output distribution (AI-8), and a distributional outlier in activation space (AI-6). The attack is at inference time (AI-2 normal) and is not hardware-related (AI-4 normal). The combination of AI-7 ceiling + HFER elevation + Fiedler drop is the spectral fingerprint of adversarial perturbation (Section 5.5), confirmed by correlated anomalies in AI-8 and AI-6.

12.3.3 Signature 3: Hardware-Induced Representational Corruption

Invariant	Reading
AI-8 (Entropy)	Anomalous (unpredictable)
AI-2 (Gradient)	Anomalous (if training active)
AI-4 (Thermal)	Violation (temperature exceeds bound)
AI-6 (Drift)	Sharp shift (corrupted activations change statistics)
AI-7 (Coherence)	Erratic (energy spikes, possibly NaN)

Interpretation: AI-4 identifies the root cause: hardware thermal violation. The representational anomalies in AI-7 are *symptoms*, not causes. Without AI-4, the Coherence Monitor would report a genuine energy anomaly but could not distinguish it from a model-caused pathology. The co-occurrence of AI-4 violation with AI-7 anomaly triggers the hardware-compromise flag (Section 11.3.3), correctly attributing the failure to the hardware substrate rather than the model.

12.3.4 Signature 4: Gradual Fine-Tuning Degradation

Invariant	Reading
AI-8 (Entropy)	Slowly declining (outputs becoming more peaked)
AI-2 (Gradient)	Trending toward bounds (norms shifting)
AI-4 (Thermal)	Normal
AI-6 (Drift)	Progressive shift (activations migrating from baseline)
AI-7 (Coherence)	Energy trending toward floor, rank declining slowly

Interpretation: Fine-tuning is progressively compressing the model's representations (AI-7 energy trending down, rank declining), causing the activation distribution to drift from the pre-fine-tuning baseline (AI-6), the output distribution to narrow (AI-8 entropy declining), and the gradient dynamics to shift (AI-2 trending). No single invariant would trigger an alarm at any given moment—all are still within their respective bands. But the *correlated temporal trends* across all four model-level invariants (AI-8, AI-2, AI-6, AI-7) signal a systemic issue. This is the value of multi-invariant trend analysis: it detects slow degradation that individual invariants miss until the violation threshold is crossed.

12.3.5 Signature 5: Distributional Shift Without Geometric Degradation

Invariant	Reading
AI-8 (Entropy)	Normal
AI-2 (Gradient)	Normal
AI-4 (Thermal)	Normal
AI-6 (Drift)	Warning or violation (distribution shifted)
AI-7 (Coherence)	Normal (energy, rank, Fiedler all within band)

Interpretation: The model is receiving inputs from a shifted distribution (AI-6 detects the shift), but its representational geometry remains healthy (AI-7 normal). This is the signature of *benign domain adaptation*: the model is operating on new data but its internal processing is sound. The AI-6 warning is legitimate (the distribution has shifted, and the model should be evaluated for accuracy on the new distribution), but AI-7's normal reading provides reassurance that the model has not geometrically degraded. Without AI-7, the AI-6 warning alone would be ambiguous—is the shift causing harm or not? The geometric health confirmation resolves this ambiguity.

12.3.6 Signature 6: Dimensional Collapse Without Energy Anomaly

Invariant	Reading
AI-8 (Entropy)	Normal or slightly reduced
AI-2 (Gradient)	Normal
AI-4 (Thermal)	Normal
AI-6 (Drift)	Normal (marginal statistics unchanged)
AI-7 (Coherence)	Energy normal, **rank below floor**

Interpretation: This is the most subtle failure mode—the one that only AI-7's dual-diagnostic architecture (energy + rank) can detect. The model's representations have lost dimensionality (rank collapse) without changing their spatial smoothness (energy normal) or their marginal statistics (AI-6 normal). Output quality may appear acceptable on standard benchmarks (AI-8 normal) because the collapsed dimensions encode features that standard benchmarks do not test. This signature is characteristic of the tunnel effect (Section 4.4.4): the model passes task-specific evaluations but has lost the representational capacity needed for transfer, OOD detection, and complex reasoning. Only the effective rank floor catches this failure.

12.4 The Health Coverage Map

The five invariants, taken together, provide coverage across a comprehensive model health space:

Table 17: Health coverage map. Checkmarks indicate direct detection capability for each dimension.

Health Dimension	AI-8	AI-2	AI-4	AI-6	AI-7
Output quality	✓				
Training stability		✓			
Hardware integrity			✓		
Distributional stability				✓	
Representational geometry					✓
Adversarial robustness	✓			✓	✓
Long-term degradation		✓		✓	✓
Hardware-model interaction			✓		✓
Fine-tuning impact		✓		✓	✓
Dimensional capacity					✓
Attention coherence					✓

12.5 The Composite Health Score

12.5.1 Definition

While each invariant produces a binary compliance verdict, a composite health score provides a scalar summary of overall model health for dashboards, trend analysis, and insurance risk scoring:

$$\mathcal{H}(t) = \prod_{k \in \{8,2,4,6,7\}} \mathcal{H}_k(t) \tag{95}$$

where $\mathcal{H}_k(t) \in [0, 1]$ is the per-invariant health score:

$$\mathcal{H}_k(t) = \begin{cases} 1.0 & \text{if invariant } k \text{ is GREEN} \\ 1.0 - \alpha_{\text{warn}} \cdot w_k & \text{if invariant } k \text{ is AMBER} \\ 0.0 & \text{if invariant } k \text{ is RED} \\ \text{undefined} & \text{if invariant } k \text{ is UNKNOWN} \end{cases} \tag{96}$$

where $\alpha_{\text{warn}} \in (0, 1)$ is the warning penalty (default: 0.1) and w_k is the severity weight for invariant k (default: equal weights $w_k = 1$ for all k).

12.5.2 Properties

- $\mathcal{H}(t) = 1.0$ if and only if all five invariants are GREEN.
- $\mathcal{H}(t) = 0.0$ if any invariant is RED (the product zeroes out).
- $\mathcal{H}(t) \in (0, 1)$ when some invariants are AMBER and none are RED.
- The multiplicative structure means that *any* RED invariant is disqualifying—there is no "averaging out" of a critical failure with healthy readings elsewhere.

Honest Framing

The composite health score is a **convenience metric** for dashboards and trend analysis, not a governance primitive. Compliance is determined by the individual invariant verdicts, not by the composite score. A system with $\mathcal{H} = 0.95$ (one invariant in AMBER) is not "95% compliant"—it is compliant with a warning active. A system with $\mathcal{H} = 0.0$ (one invariant in RED) is non-conforming, regardless of the health of the other four invariants. Binary compliance cannot be reduced to a scalar without losing its governance force.

12.6 The Completion of MAI-1

With Clause AI-7, the five mandatory invariants specified in MAI-1 §7 are complete:

MAI-1 Mandatory Invariant Suite: Complete

The MAI-1 attestation interface requires conforming systems to continuously monitor and attest the following five invariants:

1. **AI-8 (Entropy Floor):** Output token entropy within certified band.
2. **AI-2 (Gradient Stability):** Gradient norms within certified bounds during training and fine-tuning.
3. **AI-4 (SRAM Thermal Integrity):** Hardware junction temperature within certified bounds.
4. **AI-6 (Distribution Drift):** Activation distribution within certified distance of the baseline.
5. **AI-7 (Structural Coherence):** Dirichlet energy within certified band, effective rank above certified floor, attention connectivity above certified threshold.

A system claiming MAI-1 conformance at Level 2 (Model State Health) **SHALL** pass all CTS-1 assertions for all five invariants. Failure of any single invariant constitutes non-conformance with the complete Level 2 specification.

This completion has a strategic implication for the Auburn Governance Stack: all documents at the Enforcement Layer (CTS-1, test vectors, conformance levels) and the Application Layer (sector-specific compliance profiles) can now reference the complete invariant suite. No further mandatory invariants are required for the initial deployment of the governance stack. Extended invariants (MoE routing stability, inference latency bounds, quantization integrity) remain optional enhancements that complement but do not replace the mandatory five.

13 Insurability: Structural Coherence as an Insurable Property

The structural coherence invariant transforms representational health from an unobservable internal property into a certified, continuously monitored, cryptographically attested quantity. This transformation has a direct commercial consequence: it makes AI representational risk *insurable*. This section develops the insurability framework, connecting the technical specification to the actuarial and underwriting infrastructure that creates market pressure for adoption.

13.1 The Insurability Gap

13.1.1 Why AI Risk Is Currently Uninsurable

Traditional insurance requires three properties that AI systems have historically lacked:

1. **Observable risk state:** The insurer must be able to observe or verify the current risk level of the insured system. For physical assets, this is straightforward (building inspections, vehicle telemetry). For AI systems, the internal state has been opaque—no standardized metric exists for "how healthy is this model right now?"

2. **Quantifiable loss trigger:** The insurer needs an objective, machine-readable event that triggers a claim. For property insurance, this is the damage event. For AI systems, "the model produced a bad output" is subjective, contested, and difficult to verify after the fact.

3. **Actuarial history:** Premium pricing requires historical data on failure frequency and severity. For AI systems, no standardized monitoring has existed to produce this data.

AI-7, integrated with the MAI-1 attestation infrastructure, closes all three gaps.

13.1.2 How AI-7 Closes the Gap

Insurance Requirement	Traditional Analogue	AI-7 Implementation
Observable risk state	Building inspection report	Continuous energy, rank, and Fiedler monitoring with cryptographically signed attestation payloads
Quantifiable loss trigger	Fire alarm activation, seismic sensor threshold	Compliance state transition GREEN/AMBER → RED, with timestamped measurement and CoRIM-sealed band parameters
Actuarial history	Claims database, IoT sensor logs	Measurement log time series (Section 11.5), state transition logs, calibration artifacts
Risk differentiation	Building construction class, fire suppression systems	Governance risk margin tier (Table 7), monitoring configuration (Table 14), margin-to-boundary metrics

13.2 The Certified Coherence Band as Insurable Property

13.2.1 The Insurance Analogy

The coherence band $[E^D_{\min}(\ell), E^D_{\max}(\ell)]$ functions as an *operating envelope*—the AI analogue of a structural load rating for a building or an operating temperature range for industrial equipment. The model is "within specification" when its energy is within the band, and "out of specification" when it is not. This binary determination is the foundation of parametric insurance:

- The **band parameters** are the "engineering specifications" of the model, sealed in the CoRIM artifact and bound to the model version.
- The **energy measurements** are the "telemetry data" that the insurer uses to verify operating compliance.
- The **state transitions** are the "loss events" that trigger claims.
- The **margin to boundary** ($E^D_{\max}(\ell) - \hat{E}^{(\ell)}_{\text{Attn}}$ or $\hat{E}^{(\ell)}_{\text{Attn}} - E^D_{\min}(\ell)$) is the "safety factor" that determines premium differentiation.

13.2.2 The Margin-to-Boundary Metric

Define the normalized margin-to-boundary at layer ℓ:

$$m^{(\ell)}(t) = \min\left(\frac{\hat{E}^{(\ell)}_{\text{Attn}}(t) - E^D_{\min}(\ell)}{E^D_{\text{median}}(\ell) - E^D_{\min}(\ell)}, \frac{E^D_{\max}(\ell) - \hat{E}^{(\ell)}_{\text{Attn}}(t)}{E^D_{\max}(\ell) - E^D_{\text{median}}(\ell)}\right) \tag{97}$$

This metric has the following properties:

- $m^{(\ell)} = 1.0$ when the energy is at the calibration median (maximum margin).

- $m^{(\ell)} = 0.0$ when the energy is at either band boundary (zero margin—RED transition imminent).
- $m^{(\ell)} < 0$ when the energy is outside the band (violation).
- The minimum across layers $m(t) = \min_{\ell \in \mathcal{M}} m^{(\ell)}(t)$ gives the system's overall margin.

The time-averaged margin $\bar{m} = \frac{1}{W} \sum_t m(t)$ over a policy period provides the actuarial risk score: systems that consistently operate near the median ($\bar{m} \approx 1$) are lower risk than those that frequently approach the boundary ($\bar{m} \approx 0.2$).

13.3 Parametric Insurance Structure

13.3.1 Product Definition

AI-7 enables a parametric insurance product structured as follows:

Table 18: Parametric AI coherence insurance product structure.

Component	Specification
Insured event	Compliance state enters RED at any monitored layer, sustained for $\geq W_{\text{trigger}}$ consecutive measurement points (default: $W_{\text{trigger}} = 3$, preventing single-measurement false triggers).
Trigger mechanism	Binary, deterministic: the composite predicate $\mathcal{SC}(\ell, t) = 0$ for any $\ell \in \mathcal{M}$ for $\geq W_{\text{trigger}}$ consecutive t. No subjective assessment required.
Severity classes	Four classes based on violation type (Section 10.5.2): Class 1 (floor/collapse), Class 2 (ceiling/fragmentation), Class 3 (rank/dimensional), Class 4 (connectivity/attention).
Payout formula	Base payout P_{base} scaled by severity magnitude and duration: $P = P_{\text{base}} \cdot \sigma_{\text{class}} \cdot (1 + \delta \cdot D_{\text{RED}})$ where σ_{class} is the severity-class multiplier, δ is the duration scaling factor, and D_{RED} is the number of consecutive RED measurement points.
Evidence package	Machine-readable attestation payloads covering the violation period, CoRIM artifact binding the band parameters to the model version, and state transition logs with timestamps.
Exclusions	Hardware-caused violations (AI-4 co-violation), operator-induced violations (intentional model modification without recalibration), and monitor failures (UNKNOWN state).

13.3.2 Severity Class Multipliers

Table 19: Severity class definitions and payout multipliers.

Class	Violation Type	σ_{class}	Rationale
1	Energy floor violation	1.0	Collapse is progressive and predictable; remediation (recalibration, rollback) is typically feasible.
2	Energy ceiling violation	1.5	Fragmentation may indicate adversarial activity or chaotic dynamics; impact on outputs is more severe and less predictable.
3	Rank floor violation	1.2	Dimensional collapse is subtle and may have persisted undetected before AI-7 adoption; accumulated impact may be significant.
4	Connectivity violation	1.8	Attention fragmentation directly impacts reasoning coherence; highest risk of output-level harm (hallucination, incoherent reasoning).

13.3.3 Premium Differentiation

Insurance premiums are differentiated based on observable risk characteristics:

Table 20: Premium differentiation factors for AI coherence insurance.

Factor	Impact on Premium	Measurement
Governance risk margin tier	Lower premium for wider margins	Table 7: safety-critical (15%) vs. low-risk (0%)
Monitoring configuration	Lower premium for higher intensity	Table 14: Full vs. Lite
Historical margin-to-boundary	Lower premium for higher $\bar{m}$	Time-averaged margin over prior policy period
Cross-invariant monitoring	Lower premium if all five invariants monitored	MAI-1 Level 2 full conformance
Recalibration compliance	Lower premium for timely recalibration	CoRIM artifact history showing recalibration within 90-day window
Model architecture	Risk-adjusted by architecture class	Deeper models (higher collapse risk) vs. wider models (lower collapse risk)

13.4 The Actuarial Data Pipeline

13.4.1 From Monitoring to Actuarial Tables

The AI-7 monitoring infrastructure produces the raw data that actuaries need to build loss models. The pipeline from monitoring to actuarial pricing operates as follows:

1. **Measurement collection:** The Coherence Monitor produces timestamped energy, rank, and Fiedler measurements at the configured monitoring rate (Section 7.8).
2. **State transition logging:** Every compliance state transition is logged with the triggering condition, measurement values, and model version (Section 11.5).
3. **Aggregation:** Over a policy period (typically quarterly or annually), the measurement and transition logs are aggregated into summary statistics: violation frequency, violation duration distribution, severity class distribution, margin-to-boundary distribution.
4. **Loss modeling:** Actuaries use the aggregated data to estimate loss frequency (how often do violations occur?) and loss severity (how severe and how long are violations?). Standard actuarial techniques—Poisson frequency models, Pareto severity models—apply directly.
5. **Premium calculation:** The expected loss $\mathbb{E}[\text{Loss}] = \text{Frequency} \times \text{Severity}$ is loaded with expense and profit margins to produce the premium.

13.4.2 The Cold Start Problem

AI coherence insurance faces a cold start problem: before widespread AI-7 adoption, there is insufficient historical data for actuarial pricing. The resolution follows the pattern established by cyber insurance in the early 2010s:

- **Phase 1 (Current):** Expert-judgment pricing based on calibration data and stress testing. Insurers use the calibration procedure (Section 8) to establish the healthy operating range and the stress-test results to estimate violation probability under adverse conditions.
- **Phase 2 (12–24 months):** Experience-rated pricing as monitoring data accumulates across the insured portfolio. Early adopters generate the loss data that informs portfolio-wide pricing.
- **Phase 3 (24+ months):** Credibility-weighted pricing combining portfolio experience with individual policyholder monitoring data. The margin-to-boundary time series becomes the primary individual risk factor.

13.5 Market Pressure Dynamics

13.5.1 The Insurance-Procurement Feedback Loop

The insurability of structural coherence creates a self-reinforcing adoption cycle:

1. **Insurers offer coherence-conditional coverage:** AI liability policies offer lower premiums for systems with MAI-1 Level 2 conformance (all five invariants monitored).
2. **Enterprises demand conformance for procurement:** To qualify for lower insurance premiums, enterprises require their AI vendors to provide MAI-1 Level 2 conformance evidence.
3. **AI vendors adopt the Auburn stack:** To win enterprise contracts, AI vendors implement the five invariants and the attestation infrastructure.

4. **Monitoring data improves actuarial models:** As adoption increases, the volume of monitoring data grows, enabling more accurate actuarial pricing.

5. **More accurate pricing drives further adoption:** Better pricing makes insurance more attractive, expanding the insured base and strengthening the procurement pressure.

This is the same feedback loop that drove PCI-DSS adoption in payment processing and SOC 2 adoption in cloud services. The Auburn Governance Stack is designed to initiate this loop.

13.5.2 The Reinsurance Layer

Primary insurers writing AI coherence policies will seek reinsurance to manage portfolio-level risk. Reinsurers (Munich Re, Swiss Re, Hannover Re) require standardized risk metrics across the portfolio. The MAI-1 attestation format provides this standardization: every insured system reports the same fields (Table 6) in the same format, enabling portfolio-level aggregation that is impossible with bespoke monitoring approaches.

The reinsurance layer amplifies the adoption pressure: reinsurers set portfolio standards that primary insurers must meet, which flow through to policyholder requirements, which flow through to vendor specifications. Each layer of the insurance chain reinforces the standardization mandate.

13.6 The Armilla–Munich Re Pipeline

13.6.1 Existing Market Infrastructure

The AI insurance market is not hypothetical. Armilla AI, in partnership with Munich Re, has launched AI warranty and insurance products that provide coverage for AI system performance. Their current approach relies on output-level performance benchmarks (accuracy, fairness metrics) evaluated at deployment time.

AI-7 extends this approach from output-level *snapshots* to representation-level *continuous monitoring*:

Dimension	**Current (Output-Level)**	**AI-7 Extension (Representation-Level)**
Evaluation timing	Pre-deployment snapshot	Continuous runtime monitoring
Metric type	Output accuracy, fairness	Internal geometric health
Failure detection	After output-level degradation observed	Before output-level degradation manifests
Evidence standard	Benchmark test results	Cryptographically signed attestation payloads
Claims adjudication	Subjective (did the output "fail"?)	Objective (did energy exit the certified band?)

The value proposition for insurers is the shift from *reactive* coverage (pay after the bad output) to *predictive* coverage (detect the geometric precursor before the bad output). This shift reduces loss severity (earlier detection enables faster remediation) and enables more accurate pricing (continuous monitoring provides richer risk data than pre-deployment snapshots).

13.7 Summary: The Commercial Moat

The insurability framework creates a commercial moat for the Auburn Governance Stack through three mechanisms:

1. **Standardization lock-in:** Once insurers build actuarial models on MAI-1 attestation data, switching to a different attestation format requires rebuilding the entire actuarial infrastructure. The first-mover advantage in actuarial data is durable.

2. **Procurement pressure:** Insurance-conditional procurement requirements create demand that individual vendors cannot satisfy with proprietary monitoring. Only a standardized, interoperable specification works across the supply chain.

3. **Regulatory alignment:** As regulators adopt the insurance industry's risk quantification frameworks (the EU AI Act explicitly references insurance and liability), the Auburn stack's alignment with insurance requirements becomes a regulatory advantage.

Honest Framing

The insurance products described in this section do not yet exist in the specific parametric form outlined above. The Armilla/Munich Re partnership covers AI performance risk but does not yet use Dirichlet energy or representational geometry as underwriting criteria. This section describes the *architecture* of insurance products that AI-7 enables and the market dynamics that would drive their adoption. The transition from architecture to market requires insurer engagement, actuarial development, and regulatory acceptance—none of which are guaranteed. The honest framing applies: AI-7 provides the technical infrastructure for insurability; the commercial realization depends on market adoption.

14 Limitations, Open Problems, and Conclusion

The honest framing principle requires that this specification acknowledge what it cannot do as clearly as it states what it can. This section enumerates the limitations of the structural coherence invariant, identifies open research problems, and concludes with the document's contribution to the Auburn Governance Stack.

14.1 Fundamental Limitations

14.1.1 Limitation 1: Geometry Is Necessary but Not Sufficient

The structural coherence invariant certifies that the model's representational geometry is within a healthy regime. It does *not* certify that the model's outputs are correct, safe, or aligned with human values. A model with perfectly healthy geometry can still produce harmful, biased, or factually incorrect outputs if its training data, reward signal, or task specification is flawed.

The analogy is precise: a structural engineering inspection certifies that a building's load-bearing elements are within specification. It does not certify that the building is well-designed, aesthetically pleasing, or fit for its intended purpose. Structural health is a *precondition* for safe operation, not a *guarantee* of it.

14.1.2 Limitation 2: Architecture-Specific Calibration

The coherence band is calibrated empirically for a specific model architecture, training state, and deployment distribution. The calibrated parameters do *not* transfer across architectures:

- A band calibrated for a 7B-parameter decoder-only transformer is invalid for a 70B-parameter model, an encoder-decoder model, a mixture-of-experts model, or a state-space model.
- A band calibrated on English-language data is not validated for multilingual deployment without recalibration.
- A band calibrated for a pre-trained base model is invalidated by fine-tuning, quantization, or any weight modification (Section 8.7).

This is not a deficiency of AI-7 but a fundamental property of empirical calibration. Universal thresholds do not exist for representational health, just as universal load ratings do not exist for structural engineering—every structure must be evaluated individually.

14.1.3 Limitation 3: The Attention Graph Approximation

The Dirichlet energy framework was developed for message-passing on fixed graphs (GNNs). Its adaptation to transformers via the attention-as-dynamic-graph formulation (Section 3.3) is a principled but imperfect bridge:

- The attention graph is *directed* (token i attending to token j with weight A_{ij} does not imply j attends to i with the same weight). The symmetrization $\tilde{A} = (A + A^\top)/2$ loses directional information.
- The attention graph is *dynamic*—it changes with every input and every layer. The GNN theory assumes a fixed graph. The energy bounds derived for fixed graphs (e.g., the exponential decay rates of Section 4.2) are approximations when applied to dynamic attention graphs.
- The attention graph is *input-dependent*. The coherence band is calibrated from the energy distribution across a validation set, but adversarial inputs can produce attention graphs with radically different spectral properties than any validation input. The band may not fully characterize the healthy range for all possible attention topologies.

These approximations do not invalidate the framework—the empirical evidence (Sections 4–5) confirms that the adapted Dirichlet energy captures meaningful representational properties. But they do mean that the theoretical guarantees from the GNN literature apply to transformers only approximately.

14.1.4 Limitation 4: The Effective Rank Approximation

The randomized SVD (Algorithm 4, Section 9.3) provides an approximate effective rank. The approximation error depends on the spectral decay of the hidden state matrix and the chosen target rank k. For models with slowly decaying singular value spectra, the approximation may miss contributions from small singular values that collectively represent significant representational capacity.

Additionally, the effective rank is a *global* measure of dimensionality. It does not capture *local* variations in intrinsic dimensionality across the representation manifold. A model could have high global effective rank while exhibiting local dimensional collapse in specific regions of the representation space (e.g., for specific input types or token positions). The Local Intrinsic Dimensionality (LID) analysis of Ma et al. (2018) captures local structure but at significantly higher computational cost.

14.1.5 Limitation 5: Temporal Resolution

The monitoring frequency (Table 4) introduces a temporal resolution limit. At 1% monitoring rate, 99 out of every 100 requests are unmonitored. A transient coherence violation lasting a single request may go undetected. For safety-critical deployments, 100% monitoring eliminates this gap but at higher computational cost.

Even at 100% monitoring, the asynchronous computation pipeline (Section 9.5) means the coherence state is determined *after* the response is returned. A model that enters RED produces its degraded output before the violation is detected. Pre-response coherence verification (placing the energy computation on the critical path) addresses this but adds latency.

14.2 Open Research Problems

14.2.1 Problem 1: Universal Energy Normalization

The per-token normalization $\hat{E}^{(\ell)}_{\text{Attn}} = E^{(\ell)}_{\text{Attn}}/T$ provides approximate length independence, but systematic length dependence can persist (Section 8.4). A theoretically grounded normalization that provably eliminates length dependence—analogous to the batch normalization of activations—remains an open problem.

14.2.2 Problem 2: Causal Attention Energy Theory

The Dirichlet energy theory (oversmoothing rates, spectral gap tradeoffs, curvature analysis) was developed for *undirected* graphs with *symmetric* message passing. Causal (autoregressive) attention is fundamentally asymmetric: token i can attend to token $j < i$ but not vice versa. A rigorous energy theory for directed, causal attention graphs—with decay bounds, spectral characterization, and convergence results specific to the triangular attention structure—does not yet exist.

14.2.3 Problem 3: Mixture-of-Experts Adaptation

Mixture-of-Experts (MoE) architectures route different tokens through different expert networks. The representational geometry of MoE models involves not only the hidden state space but also the *routing* space—which expert processes which token. The Dirichlet energy framework does not currently account for routing-induced representational fragmentation, where tokens processed by different experts may inhabit disconnected regions of the representation space by design rather than pathology. Adapting AI-7 to MoE architectures requires distinguishing healthy expert specialization from pathological representational fragmentation.

14.2.4 Problem 4: State-Space Model Adaptation

State-space models (Mamba, S4, and successors) do not use attention matrices. They process sequences through learned linear state transitions. The attention-as-dynamic-graph formulation does not apply. Adapting the structural coherence invariant to state-space architectures requires an alternative graph construction—potentially based on the state transition matrix or the selectivity mechanism—and corresponding energy definitions.

14.2.5 Problem 5: Optimal Layer Selection

The mandatory monitored layer set $\mathcal{M} = \{L/4, L/2, 3L/4, L-1\}$ is based on empirical observations about where collapse initiates and how intrinsic dimensionality evolves. An information-theoretic framework for *optimal* layer selection—identifying the minimum set of layers that maximizes failure detection probability—remains open. The optimal set likely depends on the model architecture, training stage, and deployment distribution.

14.2.6 Problem 6: Conformal Calibration Under Distribution Shift

The conformal coverage guarantee (Section 8.3.2) assumes exchangeability between calibration and deployment data. When the deployment distribution shifts (detected by AI-6), the conformal guarantee degrades. Adaptive conformal methods (Gibbs & Candès, 2021) provide coverage guarantees under distribution shift, but their application to the multi-dimensional, multi-layer coherence band has not been developed.

14.2.7 Problem 7: Formal Verification of the Monitor

The Coherence Monitor is itself a software component that could contain bugs. Formal verification of the monitor's implementation—proving that the state machine transitions are correct, that the energy computation matches the mathematical definition, and that the Frobenius proxy screening does not produce false negatives—would strengthen the governance assurance. This verification could potentially be performed using the zero-admits Coq methodology established elsewhere in the Auburn Patent Family.

14.3 What This Document Provides

Despite the limitations enumerated above, Clause AI-7 provides:

1. **The first governance-grade specification for representational health monitoring.** No prior work has translated the academic literature on oversmoothing, oversquashing, neural collapse, and dimensional collapse into a deployable, binary-compliance, continuously-monitored invariant with cryptographic attestation.

2. **A dual-diagnostic framework** (Dirichlet energy + effective rank) that provides strictly more comprehensive coverage than either metric alone, grounded in the theoretical result of Zhang et al. (2025) that rank collapse is strictly more general than energy collapse.

3. **A computationally tractable implementation** achieving $< 2\%$ latency overhead through six orthogonal optimizations, demonstrated across model sizes from 7B to 405B parameters.

4. **A calibration methodology** with distribution-free conformal coverage guarantees, governance risk margins, and a recalibration protocol that maintains the guarantee across model updates.

5. **A regulatory mapping** demonstrating compliance with five regulatory frameworks (EU AI Act, SR 11-7, FDA PCCP, NIST AI RMF, Solvency II) that require continuous monitoring but prescribe no specific internal metrics.

6. **The completion of the MAI-1 mandatory invariant suite.** With AI-7, all five invariants are specified, enabling the full Layer 2 conformance assessment and unlocking the Enforcement and Application layers of the Auburn Governance Stack.

7. **An insurability framework** that transforms representational health from an unobservable property into a continuously monitored, cryptographically attested, parametrically insurable risk metric.

14.4 Conclusion

Clause AI-7 answers a question that the AI governance community has not previously asked in operational terms: *Is the model's representational geometry healthy right now?*

The question matters because representational health is the geometric precondition for everything else a model does. A model with collapsed representations cannot distinguish between

inputs. A model with fragmented representations cannot reason coherently across its context. A model with dimensional degeneration has lost capacity that no output-level evaluation can detect. These are not theoretical concerns—they are documented, empirically characterized pathologies that affect production systems.

The contribution of this specification is to make representational health *observable, measurable, certifiable, and insurable.* The Dirichlet energy band provides the primary invariant. The effective rank floor closes the gap that energy alone cannot cover. The Fiedler connectivity threshold catches attention fragmentation. The compliance state machine provides deterministic, auditable governance. The calibration procedure provides distribution-free coverage guarantees. The computational optimizations make monitoring feasible at production scale. The regulatory mapping demonstrates immediate applicability. The cross-invariant integration completes the five-invariant suite. The insurability framework creates the market pressure for adoption.

The structural coherence invariant does not guarantee that a model is safe, accurate, or aligned. It guarantees that a model's internal geometry has not entered a regime that is *known to precede* safety, accuracy, and alignment failures. This is the honest framing: not a safety guarantee, but a certified absence of known geometric pathology. In a governance landscape where no internal monitoring exists, this is the necessary first step.

The Auburn Governance Stack's Layer 2 is now complete. The five mandatory invariants—entropy, gradient stability, thermal integrity, distributional drift, and structural coherence—provide comprehensive coverage of the model health space. Every document at the Enforcement and Application layers can now reference the complete invariant suite. The infrastructure specification is ready for deployment.

References

[1] Li, Q., Han, Z., and Wu, X.-M. (2018). "Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). The foundational identification of oversmoothing in graph convolutional networks.

[2] Oono, K. and Suzuki, T. (2020). "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification." *International Conference on Learning Representations (ICLR 2020)*. Rigorous proof of exponential convergence to collapsed equilibrium in GNNs.

[3] Dong, Y., Cordonnier, J.-B., and Loukas, A. (2021). "Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth." *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*. Doubly exponential rank collapse result for pure self-attention.

[4] Cai, C. and Wang, Y. (2020). "A Note on Over-Smoothing for Graph Neural Networks." *arXiv preprint arXiv:2006.13318*. Discrete Dirichlet energy decay bounds for GNNs.

[5] Rusch, T.K., Bronstein, M.M., and Mishra, S. (2023). "A Survey on Oversmoothing in Graph Neural Networks." *arXiv preprint arXiv:2303.10993*. Comprehensive survey establishing that constant energy is necessary but not sufficient for deep GNN performance.

[6] Di Giovanni, F., Rowbottom, J., Chamberlain, B.P., Markovich, T., and Bronstein, M.M. (2023). "Understanding Convolution on Graphs via Energies." *Transactions on Machine Learning Research (TMLR)*. Proof that graph convolutions can sharpen (increase energy) when weight matrices have negative eigenvalues.

[7] Di Giovanni, F., Giusti, L., Barbero, F., Luise, G., Liò, P., and Bronstein, M.M. (2023). "On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology." *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Relative importance of width, depth, and topology for oversquashing.

[8] Alon, U. and Yahav, E. (2021). "On the Bottleneck of Graph Neural Networks and its Practical Implications." *International Conference on Learning Representations (ICLR 2021)*. Identification of oversquashing as a fundamental GNN limitation with exponential gradient decay through bottlenecks.

[9] Topping, J., Di Giovanni, F., Chamberlain, B.P., Dong, X., and Bronstein, M.M. (2022). "Understanding Over-Squashing and Bottlenecks on Graphs via Curvature." *International Conference on Learning Representations (ICLR 2022)*. Connection of oversquashing to negative Ricci curvature; Balanced Forman Curvature as a tractable diagnostic.

[10] Black, M., Wan, Z., Nayyeri, A., and Wang, Y. (2023). "Understanding Oversquashing in GNNs through the Lens of Effective Resistance." *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Total effective resistance as a global oversquashing measure.

[11] Nguyen, K. et al. (2023). "Revisiting Over-smoothing and Over-squashing Using Ollivier-Ricci Curvature." *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Unified curvature framework connecting both failure modes.

[12] Bronstein, M.M., Bruna, J., Cohen, T., and Veličković, P. (2021). "Geometric Deep Learning: Grids, Groups, Graphs, and Geodesics." *arXiv preprint arXiv:2104.13478*. Comprehensive framework for geometric deep learning; graph Fourier basis definition.

[13] Papyan, V., Han, X.Y., and Donoho, D.L. (2020). "Prevalence of Neural Collapse during the Terminal Phase of Deep Learning Training." *Proceedings of the National Academy of Sciences*, 117(40), 24652–24663. Discovery and documentation of the four NC properties (NC1–NC4).

[14] Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. (2021). "A Geometric Analysis of Neural Collapse with Unconstrained Features." *Advances in Neural Information Processing Systems (NeurIPS 2021)*. Proof that Simplex ETF is the unique global minimizer of cross-entropy with weight decay.

[15] Wu, Y. and Papyan, V. (2024). "Neural Collapse in Large Language Models." *arXiv preprint*. Documentation of linguistic neural collapse phenomena in autoregressive LLMs.

[16] Jing, L., Vincent, P., LeCun, Y., and Tian, Y. (2022). "Understanding Dimensional Collapse in Contrastive Self-Supervised Learning." *International Conference on Learning Representations (ICLR 2022)*. Root causes of dimensional collapse in self-supervised learning.

[17] Masarczyk, W., Ostaszewski, M., Imani, E., Pascanu, R., Miłoś, P., and Trzcinski, T. (2023). "The Tunnel Effect: Building Data Representations in Deep Neural Networks." *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Discovery of the tunnel effect: deep layers compressing representations to task-minimum rank.

[18] Ansuini, A., Laio, A., Macke, J.H., and Zoccolan, D. (2019). "Intrinsic Dimension of Data Representations in Deep Neural Networks." *Advances in Neural Information Processing Systems (NeurIPS 2019)*. The "hunchback" intrinsic dimensionality profile across CNN layers.

[19] Valeriani, L. et al. (2023). "The Geometry of Hidden Representations of Large Transformer Models." *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Intrinsic dimensionality profiles in large transformers (ESM-2, iGPT).

[20] Feng, S., Zheng, Y., Huang, W., Zhao, P., Jordan, M.I., and Zha, H. (2022). "Rank Diminishing in Deep Neural Networks." *Advances in Neural Information Processing Systems (NeurIPS 2022)*. Universal monotone decreasing property of network rank under composition.

[21] Daneshmand, H., Kohler, J., Bach, F., Hofmann, T., and Lucchi, A. (2020). "Batch Normalization Provably Avoids Ranks Collapse for Randomly Initialised Deep Networks." *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Proof that batch normalization preserves rank $\Omega(\sqrt{d})$.

[22] Zhang, Y., Wei, C., Xu, Z., and Liu, Z. (2025). "On the Relationship between Rank Collapse and Energy Collapse in Graph Neural Networks." *International Conference on Learning Representations (ICLR 2025)*. Proof that rank collapse is strictly more general than energy collapse.

[23] Dohare, S., Lan, J.Q., and Sutton, R.S. (2024). "Loss of Plasticity in Deep Continual Learning." *Nature*, 632, 768–774. Documentation of loss of plasticity in continual learning with Type-1/Type-2 characterization.

[24] Roy, O. and Vetterli, M. (2007). "The Effective Rank: A Measure of Effective Dimensionality." *15th European Signal Processing Conference (EUSIPCO 2007)*. Definition of effective rank via Shannon entropy of normalized singular values.

[25] Halko, N., Martinsson, P.G., and Tropp, J.A. (2011). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." *SIAM Review*, 53(2), 217–288. The foundational randomized SVD algorithm with error bounds.

[26] Ma, X., Li, B., Wang, Y., Erfani, S.M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M.E., and Bailey, J. (2018). "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality." *International Conference on Learning Representations (ICLR 2018)*. Elevated LID in adversarial regions; LID-based adversarial detection with AUC > 0.90.

[27] Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. (2019). "Adversarial Robustness as a Prior for Learned Representations." *arXiv preprint arXiv:1906.00945*. Adversarially trained models produce smoother, approximately invertible representations.

[28] Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. (2020). "Do Adversarially Robust ImageNet Models Transfer Better?" *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Robust ImageNet models transfer better to 12 downstream tasks.

[29] Cheng, Y., Zhu, M., Zhang, D., and Liu, S. (2022). "Feature Spectral Regularization for Robustness." *arXiv preprint*. Low-eigenvalue eigenvectors are more non-robust; spectral regularization improves adversarial robustness.

[30] Khachaturov, D. et al. (2024). "Effective Dimensionality and Adversarial Robustness." *arXiv preprint*. Near-linear inverse relationship between effective dimensionality and adversarial robustness.

[31] Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. (2017). "Parseval Networks: Improving Robustness to Adversarial Examples." *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*. Lipschitz-constrained weight matrices as approximate Parseval tight frames for robustness.

[32] Shi, H. et al. (2022). "Revisiting Over-smoothing in BERT from the Perspective of Graph." *arXiv preprint arXiv:2202.08625*. Analysis of transformer oversmoothing from the graph perspective; role of LayerNorm.

[33] Godey, N. et al. (2024). "Anisotropy Is Inherent to Self-Attention in Transformers." *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*. Proof that anisotropic clustering is an inherent property of self-attention.

[34] Guo, X. et al. (2023). "Contrastive Learning for Non-Local Graphs with Multi-Resolution Structural Views (NeuTRENO)." *arXiv preprint*. Bridge between GNN oversmoothing and transformer oversmoothing.

[35] Zhai, S. et al. (2023). "Stabilizing Transformer Training by Preventing Attention Entropy Collapse." *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Attention entropy bounds and their connection to oversmoothing.

[36] Rusch, T.K., Chamberlain, B.P., Rowbottom, J., Mishra, S., and Bronstein, M.M. (2022). "Graph-Coupled Oscillator Networks." *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*. Second-order ODE framework where zero-energy steady states are not exponentially stable.

[37] Bison, F. et al. (2024). "On the Expressive Power of Spectral Invariant Graph Neural Networks." *arXiv preprint*. Demonstration that normalized Dirichlet energy can reach zero even when representations are not fully identical.

[38] Lyle, C., Zheng, Z., Nikishin, E., Pires, B.A., Pascanu, R., and Dabney, W. (2023). "Understanding Plasticity in Neural Networks." *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Effective rank as a correlate of network plasticity.

[39] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer. Foundational text on conformal prediction with finite-sample coverage guarantees.

[40] Gibbs, I. and Candès, E. (2021). "Adaptive Conformal Inference Under Distribution Shift." *Advances in Neural Information Processing Systems (NeurIPS 2021)*. Conformal methods with coverage guarantees under distribution shift.

[41] European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

[42] Board of Governors of the Federal Reserve System (2011). *SR 11-7: Supervisory Guidance on Model Risk Management*. Federal Reserve.

[43] Office of the Comptroller of the Currency (2011). *OCC 2011-12: Sound Practices for Model Risk Management*. OCC Bulletin.

[44] U.S. Food and Drug Administration (2023). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions*. FDA Guidance Document.

[45] National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1.

[46] European Parliament and Council of the European Union (2009). *Directive 2009/138/EC on the Taking-Up and Pursuit of the Business of Insurance and Reinsurance (Solvency II)*. Official Journal of the European Union.

[47] Fields, R. (2026). "The Model Attestation Interface (MAI-1): A Normative Profile and Conformance Protocol for Foundation Model Governance." *Auburn Patent Family*, Clause AI-5. Figshare.

[48] Fields, R. (2026). "CTS-1: MAI-1 Conformance Test Suite." *Auburn Patent Family*. Figshare.

[49] Fields, R. (2026). "The Model State Attestation Framework (MSAF): Three-Tier Architecture." *Auburn Patent Family*. Figshare.

[50] Fields, R. (2026). "Clause AI-6: Distribution Drift Bound." *Auburn Patent Family*. Figshare.

[51] Fields, R. (2026). "Clause AI-8: Entropy Collapse Constraint." *Auburn Patent Family*. Figshare.

[52] Fields, R. (2026). "Clause AI-2: Gradient Starvation Envelope." *Auburn Patent Family*. Figshare.

[53] Fields, R. (2026). "Clause AI-4: SRAM Thermal Integrity Bound." *Auburn Patent Family*. Figshare.

[54] Fields, R. (2026). "Rails Symposium: Capstone Tutorial for the Auburn Governance Stack." *Auburn Patent Family*. Figshare.

[55] Fields, R. (2026). "Auburn Governance Stack Master Architecture Plan (AGS-1)." *Auburn Patent Family*. Private specification.

Intellectual Property Declaration

Auburn Patent Family Fields

Intellectual Property (IP) Declaration

The methods, logic structures, and "Certified Constant" registries contained in this work are the sole property of **Ryan Fields**.

Public License (Non-Commercial)

This work is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)** license.

- **Academic Use:** Researchers may share and use this framework for non-commercial academic purposes, provided full attribution is given to Ryan Fields.
- **No Derivatives:** No modifications or "remixes" of the "Certified Constants" or logical proofs are permitted without express written consent.

Commercial Prohibition

Commercial use of this framework is strictly prohibited. This includes, but is not limited to:

- Use within proprietary high-frequency trading (HFT) risk models.
- Integration into commercial high-assurance AI governance software.
- Use by private financial institutions for "tail-risk" auditing of prime distribution variance.
- Integration into commercial AI monitoring, attestation, or compliance platforms.
- Incorporation into insurance underwriting models or actuarial pricing engines.

UncleBroFields@proton.me
fieldsryanchristopher@gmail.com

Clause AI-7: Structural Coherence Bound
Auburn Patent Family Fields
Ryan Fields, 2026

Published on Figshare under CC BY-NC-ND 4.0