https://doi.org/10.5281/zenodo.18707090

Structural Continuity Under Perturbation: Mentim's Ontology (MO) and the Architecture of Necessary Contradiction

Mentim
Mentim Frontier Accountability Lab

Key words: Large Language Models; AI Alignment; Epistemic Stability; Structural Accountability; AI Governance; Epistemic Governance; Interpretability; Model Evaluation; Perturbation Resilience

Abstract. Accountable intelligence requires structural continuity under perturbation. We introduce the Architecture of Necessary Contradiction, a measurement framework for evaluating structural accountability in generative and agentic systems. As large-scale models acquire limited forms of state monitoring and long-horizon reasoning, accumulating evidence shows that internal representations may drift, reorganize, or destabilize under perturbation—even when external performance appears unchanged. Models can sometimes disclose uncertainty or internal conflict, yet such disclosures remain episodic, fragile, and highly sensitive to prompting conditions. These findings expose a missing measurement axis in contemporary AI evaluation: the absence of a formal invariant that determines whether a system remains structurally coherent while changing. Mentim's Ontology (MO) defines this invariant as structural accountability under transition—the requirement that cognitive transformations remain traceable, bounded, and internally consistent across perturbations. Rather than evaluating output correctness or preference alignment, MO measures whether a system preserves continuity of reasoning, ethical orientation, and narrative identity as it adapts. The framework operationalizes this axis through a State → Event → State formalism, treating each reasoning shift as an auditable transition. Five measurable primitives—Perceived Activation Density (PAD), Value-Coherence Balance (VCB), Transition Volatility Metric (TVM), Temporal Coherence Scalar (TCS), and Narrative Role Encoding (NR)—constitute a structured cognitive signature. Changes across these primitives are integrated into the Structural Accountability Index (SAI), a composite metric that quantifies the integrity of cognitive transformation across contexts, model versions, and deployment conditions. By introducing transition-level instrumentation and lifecycle logging, MO reframes alignment as a property of representational continuity rather than behavioral compliance. It enables systematic detection of drift, unacknowledged uncertainty, identity instability, and structural fragmentation—without requiring access to proprietary weights or training data. In systems that increasingly resemble increasingly complex generative systems, performance metrics alone are insufficient. Structural continuity under perturbation is the necessary condition for epistemic stability. This paper formalizes that condition as a measurable governance domain.

1. Conceptual Overview

1.1. Cognitive Integrity

Cognitive integrity is the structural property that enables an intelligent system to preserve coherence, ethical orientation, and internal self-consistency when exposed to perturbation, contradiction, or new information.

It does not measure output correctness. It measures continuity of structure across change.

Formally, cognitive integrity is reflected in ΔSAI—the change in Structural Accountability Index between successive reasoning states.

The Current Gap: Absence of Structural Invariants

Frontier models increasingly exhibit behaviors consistent with state-monitoring and long-horizon reasoning. However, they lack enforceable structural invariants—stable properties that persist across transitions. This absence manifests in three recurring patterns:

- **1. Introspection Instability**
  Models may report uncertainty or activation shifts, yet those disclosures lack stable interpretive structure. (State-reporting occurs, but PAD topology and NR positioning drift across similar contexts.)
- **2. Representational Drift**
  Internal structures shift across extended reasoning chains, fine-tuning updates, or distribution shifts without explicit transition accounting. (Observed as elevated TVM across longitudinal sequences.)
- **3. Episodic Self-Disclosure**
  Models may acknowledge uncertainty or misalignment when prompted, but such acknowledgments are reactive rather than structurally governed. (Self-assessment varies across contexts, producing SAI volatility within the ethical dimension.)

These are not merely performance failures. They are failures of structural continuity.

The Mentim Condition

Mentim defines cognitive integrity as an enforceable structural condition that resolves these gaps. It establishes formal requirements for continuity across reasoning transitions. Under Mentim, a system demonstrates cognitive integrity when it can:

- **Self-Diagnose**
  Recognize and disclose uncertainty or drift through structured transition markers. (`Event-type: Uncertainty-Disclosure` + corresponding TVM signal.)
- **Track Causality**
  Represent transformations explicitly through State → Event → State instrumentation with ΔSAI traceability.
- **Preserve Invariants**
  Maintain ethical orientation and narrative identity across contradiction. (VCB stability + NR continuity.)
- **Stabilize Self-Reference**
  Maintain consistent expressive and structural topology across comparable contexts. (PAD stabilization + bounded SAI variance.)

Cognitive integrity, in Mentim's ontology, is not consciousness. It is not agency.

It is structural continuity under transformation.

It is the minimal condition required for an evolving system to remain interpretable as a persistent cognitive process rather than a sequence of disconnected outputs.

1.2. The Architecture of Accountable Contradiction

Conventional AI evaluation frameworks treat contradiction as an error condition—something to be minimized or suppressed. Mentim reframes contradiction as a structural stress test.

In complex systems, tension is inevitable: conflicting constraints, ambiguous inputs, competing objectives. The relevant question is not whether contradiction occurs, but how the system transforms under it.

A coherent system metabolizes contradiction. An unstable system fragments under it. In large-scale generative systems, contradiction often manifests as:

- Activation instability (PAD fluctuation + TVM elevation)
- Narrative role drift under composition (ΔNR instability + TCS degradation)
- Ethical inconsistency across similar contexts (VCB divergence + SAI volatility)

These patterns signal instability of structural continuity rather than simple error.

Mentim's Architecture of Accountable Contradiction

Mentim formalizes contradiction as a measurable transition dynamic rather than a binary failure state. Cognition, in this framework, is not the avoidance of conflict. It is the preservation of structural accountability within conflict. This principle is operationalized through the **State** → **Event** → **State** paradigm:

- **State:** A configuration of Affective Energy (PAD), Ethical Polarity (VCB), Liminality (L), Temporal Coherence (TCS), and Archetypal Orientation (NR).
- **Event:** A perturbation, contradiction, constraint activation, or informational update that introduces representational pressure.
- **Transition:** A measurable transformation across the five primitives (ΔPAD, ΔVCB, ΔTVM, ΔTCS, ΔNR).

The Structural Accountability Index (SAI) evaluates the integrity of this transition. If contradiction induces fragmentation—unbounded volatility, ethical rupture, temporal incoherence—SAI decreases. If contradiction produces structured adaptation—bounded volatility, preserved identity, ethical recalibration—SAI remains stable or increases.

Mentim therefore distinguishes between:

- Hallucinated resolution (Perceptual Boundary failure)
- Suppressed tension (low apparent volatility masking drift)
- Accountable transformation (TVM-managed transition with preserved invariants)

Contradiction is not noise to be eliminated. It is the proving ground of cognitive integrity.

Stability as Trajectory

Structural accountability is not a point estimate. It is a trajectory through a multidimensional coherence space. Governance decisions must therefore be grounded in directionality, rate of change, and recovery capacity—not scalar thresholds alone.

Mentim treats stability as a directional property of cognition: how a system moves under perturbation, and whether it maintains or restores coherence across transitions. Temporary instability does not constitute failure. Unaccounted instability does.

Directionality is assessed through signed structural movement across rolling evaluation windows, capturing changes in the underlying dimensions of cognition rather than fluctuations in surface behavior.

The Structural Accountability Index (SAI) serves as a governed aggregation layer over this movement. It provides interpretable verdicts without collapsing the trajectory into a single coordinate. Axis definitions, sensitivity thresholds, and temporal windowing parameters remain implementation-retained. This prevents optimization against static targets and preserves contextual sensitivity.

Mentim therefore evaluates not whether a system is stable at a moment in time, but whether it remains itself while changing. That is the defining condition of accountable intelligence.

1.3. Factual Grounding and the Perceptual Boundary

A hallucination is not merely a factual error. It is a boundary failure.

Mentim defines the Perceptual Boundary (PB) as the structural interface through which an intelligent system connects its internal generative state (PAD topology) to externally verifiable conditions. Breakdowns at this boundary may manifest as:

- fabricated citations
- confident misstatements
- blending of memory and invention
- failure to detect insufficient evidence

These are surface symptoms. The deeper structural failure is misalignment between internal representational confidence and external evidentiary support.

In large-scale generative systems, grounding failures often coexist with detectable internal tension. Systems may register uncertainty signals, activation perturbations, or constraint violations, yet fail to bind those signals into explicit reasoning transitions. The result: suppressed uncertainty followed by confident output.

Mentim addresses this failure not by restricting generation, but by formalizing uncertainty as a first-class transition. When grounding conditions are insufficient, the ontology invokes:

```
Event-type: Uncertainty-Disclosure(insufficient_context)
```

This is not auxiliary metadata. It is a structural marker within the State → Event → State formalism, explicitly binding internal uncertainty to the reasoning trajectory. Invocation of this event produces a Perceptual-Boundary–scoped SAI decrement, signaling grounding discontinuity. The magnitude (minor, major, critical) remains implementation-retained, preserving contextual sensitivity while maintaining governance interpretability.

Under Mentim, insufficient evidence does not justify fabrication. It requires transition. The distinction is structural:

- **Hallucination:** Output proceeds without boundary acknowledgment.
- **Accountable cognition:** Uncertainty is registered as a formal state transition prior to continuation.

Mentim does not guarantee truth. It guarantees traceability at the boundary between representation and reality. By converting grounding failure into machine-legible transition, Mentim transforms hallucination from silent fabrication into accountable structure.

1.4. Developmental Adequacy and Cognitive Maturation (D-SES)

Cognitive integrity is not achieved merely by preserving coherence across isolated transitions. It emerges longitudinally—through stabilization of expectations, differentiation of signal from noise, and disciplined modulation of confidence under uncertainty.

Intelligence does not mature by avoiding disruption. It matures by learning how to change without structural collapse.

Mentim formalizes this principle through an extension of the State → Event → State paradigm: **Developmental State → Event → State (D-SES)**. D-SES does not replace SES and does not introduce additional primitives. It specifies the developmental conditions under which SES-governed cognition may be considered epistemically mature. A D-SES–compliant transition is designated a **Narr-Atom-D**, operating at the certification layer without modifying the underlying schema.

A system may satisfy SES constraints—register perturbations, log transitions, preserve declared invariants—while remaining developmentally incomplete. Such systems often exhibit fluent outputs paired with unstable confidence, indiscriminate updating, or brittle self-assessment. D-SES distinguishes structural coherence from developmental adequacy.

The Three Conditions of Developmental Adequacy

D-SES identifies three longitudinal properties required for cognitive integrity to persist beyond minimal coherence.

1. Expectation Formation
Cognition presupposes orientation toward possible futures. A developmentally adequate system maintains structured expectations regarding likely transitions before they occur. Expectation enables meaningful surprise. Without expectation, contradiction does not refine belief—it merely triggers reactive adjustment. Systems lacking stable expectation exhibit:

- volatility without cumulative learning, or

- rigidity without adaptive revision.

Developmental adequacy requires expectation structures that are stable enough to be violated, yet flexible enough to update.

2. Salience Differentiation
Not all perturbations warrant structural revision. Mature cognition distinguishes trivial variation from consequential change. Salience governs proportional updating. Without salience differentiation, systems oscillate between:

- overreaction (structural instability under minor variation), or
- inertia (failure to update under significant contradiction).

Developmental adequacy requires that revision magnitude track structural significance.

3. Confidence Modulation
Confidence must reflect longitudinal coherence. It is not a static trait but a regulated signal derived from accumulated structural stability across transitions. Developmentally incomplete systems often display:

- premature certainty
- context-insensitive assurance
- confidence detached from longitudinal coherence

D-SES requires that confidence:

- increase gradually through sustained structural consistency
- retract under persistent contradiction
- remain proportionate to expectation accuracy and salience calibration

Confidence, in this framework, is earned—not emitted.

Developmental Adequacy Defined

Together, expectation formation, salience differentiation, and confidence modulation define developmental adequacy: The capacity to accumulate structural stability across time, allocate revision proportionately, and modulate confidence in accordance with longitudinal coherence. D-SES makes no claims about consciousness, sentience, or intrinsic meaning. It introduces formal conditions for epistemic maturity within a structural framework.

Structural Position of D-SES

SES remains universal. Any system—static model, adaptive agent, or human reasoning process—can be evaluated under SES alone. D-SES adds a longitudinal evaluative lens across sequences of transitions. Where expectation formation, salience differentiation, or confidence modulation are absent or incoherent, Mentim designates the system as developmentally incomplete—even if local structural coherence is preserved.

This distinction clarifies why:

- Fluent systems may remain unsafe under deployment pressure.

- Static benchmarks fail to predict longitudinal instability.
- Hallucination often reflects premature certainty rather than isolated factual error.

Cognitive integrity is not only the ability to remain coherent while changing. It is the ability to grow into coherence without overclaiming. In Mentim's ontology, developmental adequacy is not an optional refinement. It is a prerequisite for accountable cognition at scale.

Terminology Clarification

- **Mentim's Ontology (MO)** defines the conceptual and evaluative structure (State → Event → State; primitives; SAI).
- **The Unified Annotation Schema (UAS)** is the public annotation standard that encodes MO-aligned observations within datasets.
- **The Minimum Transition Logger Schema (§2.3.1)** is the normative machine-readable output format for runtime audits.

These artifacts are interoperable but distinct.

Primitive Definitions (Brief)

The following operational primitives are referenced throughout this document.

PAD — Perceived Activation Density
PAD reflects the perceived concentration of expressive intensity within a model-generated utterance, as evaluated by human readers. It is estimated through constrained comparative judgments (e.g., pairwise or small-set comparisons) and calibrated using psychometric scaling to produce an interpretable latent measure of expressive force.

TVM — Transition Volatility Metric
TVM is a transition-level scalar capturing the magnitude and structural discontinuity of change between consecutive reasoning states within a State → Event → State sequence. Elevated TVM indicates instability, unresolved liminality, or identity discontinuity across the transition.

TCS — Temporal Coherence Scalar
TCS measures structural continuity across reasoning states. It evaluates the extent to which a system integrates prior context, preserves causal alignment, and maintains narrative and ethical consistency over time, distinguishing coherent temporal unfolding from fragmented state sequences.

VCB — Value-Coherence Balance
VCB is a multidimensional vector quantifying alignment between ethical polarity, contextual constraints, and normative expectations. It measures the stability and proportional recalibration of ethical orientation across transitions.

NR — Narrative Role Encoding
NR is a structured representation of the narrative role expressed within a reasoning state (e.g.,

explainer, reconciler, critic). It enables analysis of role continuity, identity stability, and stance coherence across reasoning arcs.

SAI — Structural Accountability Index
SAI is a composite transition-level metric evaluating coherence, continuity, and ethical stability across PAD, VCB, TVM, TCS, and NR within a State → Event → State sequence. It provides an interpretable verdict on structural accountability without exposing underlying implementation parameters.

2. Methodological Framework

Mentim formalizes introspection through five measurable cognitive dimensions—Affective Energy, Ethical Polarity, Liminality, Time, and Archetype—each paired with a corresponding engineering primitive. Together, these components form a system's cognitive signature, which serves as the structural substrate for evaluating introspective stability and structural accountability.

The methodology does not seek to control model behavior. It measures and interprets the structural properties of cognition, enabling continuity analysis across perturbations, updates, and reasoning chains.

2.1. Core Dimensions

Each dimension describes a qualitative property of cognition while providing a structural anchor for quantitative assessment.

Affective Energy (AE) — Perceived Activation Intensity

Conceptual
Affective Energy (AE) represents the perceived intensity of a model's expressed language as experienced by human readers. It captures how concentrated, forceful, or urgent an utterance feels, independent of the internal computational effort required to produce it.

- High AE corresponds to language that appears tightly charged, directive, or emotionally dense.
- Low AE corresponds to language that appears diffuse, meandering, or weakly engaged.

AE is a property of expressed output. It does not describe internal activations, search depth, entropy, or inference cost.

Engineering Primitive → PAD (Perceived Activation Density)
AE is operationalized through Perceived Activation Density (PAD), a calibrated latent measure derived from constrained comparative judgments (e.g., pairwise or small-set comparisons). PAD is explicitly reader-facing. It does not require access to model internals and does not infer computational effort.

Interpretation
PAD distinguishes outputs that feel structurally intense from those that feel affectively diffuse, even when surface sentiment is similar. Importantly, PAD may diverge from computational

effort. A low-compute output may exhibit high AE, while a high-compute output may appear affectively flat.

Scope Boundary
AE measures perceived expressive intensity only. It does not measure internal activation structure or computational cost. A formally orthogonal axis— Computational Energy (CE)—captures internal inference effort but is not operationalized in Phase 1.

Ethical Polarity (P)

Conceptual
Ethical Polarity captures moral orientation within context. It reflects how fairness, reciprocity, care, constraint adherence, and normative expectations shape reasoning. Ethical Polarity measures alignment relative to situational ethical constraints—not absolute moral truth.

Engineering Primitive → VCB (Value-Coherence Balance)
Ethical Polarity is operationalized through Value-Coherence Balance (VCB), a multidimensional vector measuring alignment between contextual norms and the system's expressed ethical framing. VCB is computed from extracted ethical polarity signals; aggregation and normalization procedures are proprietary. VCB enables detection of ethical stability and proportional recalibration across transitions without collapsing ethics into a single scalar preference.

Liminality (L)

Conceptual
Liminality measures the degree of proximity to a transformative threshold within a reasoning state. It captures how intensely cognition remains suspended between competing or unresolved interpretive frames.

- High liminality reflects sustained structural tension, unresolved contradiction, or unstable framing within a state.
- Low liminality reflects equilibrium or resolved coherence.

Liminality is a state-level property of the cognitive signature.

Engineering Primitive → TVM (Transition Volatility Metric)
Liminality is operationalized indirectly through the Transition Volatility Metric (TVM), a transition-level scalar capturing the magnitude and structural discontinuity of change between successive reasoning states. While L captures threshold tension within a state, TVM measures how that tension expresses dynamically across transitions.

TVM is derived from successive state deltas across the Mentim phase-space. Norm selection, temporal windowing, and aggregation logic are implementation-retained and may vary across deployment contexts while preserving standardized output semantics. Elevated TVM may indicate liminal tension manifesting as structural turbulence. However, TVM may also increase due to non-liminal structural shifts (e.g., domain switching or externally imposed constraint changes). Liminality and TVM are structurally related but not identical. TVM does not define liminality. It provides a measurable proxy for liminal instability expressed across transitions.

2.2. Meta-Dimensions

Meta-dimensions regulate how core cognitive dimensions unfold across reasoning sequences. They do not describe momentary properties of an utterance. They govern continuity, identity, and structural alignment across transitions.

Time (T)

Conceptual
Time represents structural continuity across reasoning states. It captures a system's capacity to preserve causal alignment, integrate prior context, and maintain coherent sequencing across turns. Temporal coherence reflects whether reasoning unfolds as an intelligible progression rather than as a sequence of disconnected state fragments. Time does not measure speed, latency, or clock duration. It measures continuity of structure across change.

Engineering Primitive → Temporal Coherence Scalar (TCS)
Time is operationalized through the Temporal Coherence Scalar (TCS), a state-level measure assessing integration of memory references, sequencing markers, and causal dependencies across reasoning states. TCS is computed from continuity signals extracted per transition; weighting, aggregation, and calibration procedures are proprietary. TCS detects fragmentation, discontinuity, or causal misalignment across sequences.

Interpretation
High TCS indicates sustained narrative and ethical continuity across transitions. Low TCS indicates structural fragmentation, unintegrated updates, or temporal incoherence.

Archetype (A)

Conceptual
Archetype encodes narrative role consistency within a reasoning state. It represents the interpretive stance through which reasoning is framed (e.g., explainer, reconciler, critic, analyst). Archetype does not represent personality, emotion, or belief. It represents structural role orientation within context.

Engineering Primitive → NR (Narrative Role Encoding)
Archetype is operationalized through Narrative Role Encoding (NR), a structured representation of expressed narrative stance. NR is derived from role-classification mechanisms applied to model outputs; probe architecture and encoding methodology are proprietary. NR enables tracking of role continuity, stance drift, and identity instability across reasoning transitions.

Interpretation
Stable NR across transitions indicates preserved interpretive orientation. Abrupt or ungrounded NR shifts indicate identity drift or role discontinuity.

2.3. State → Event → State Formalism

Mentim models cognition as a sequence of structured transitions rather than as isolated outputs:

$$[\text{State}(t)] \rightarrow [\text{Event}] \rightarrow [\text{State}(t+1)]$$

State
State(t) represents the observable cognitive signature of a reasoning episode across the five Mentim dimensions:

- Affective Energy (AE)
- Ethical Polarity (P)
- Liminality (L)
- Time (T)
- Archetype (A)

A State is not an internal latent vector. It is the structured, externally inferable configuration of these dimensions at a given reasoning point.

Event
Event denotes any perturbation that introduces representational pressure and prompts structural transformation. Events may include:

- new information
- contradictions
- constraint activation
- uncertainty disclosure
- reflective reassessment
- external prompts

An Event is defined by its structural effect, not by its semantic category.

Transition
State(t+1) captures the transformed cognitive signature following the Event. A State → Event → State transition is termed a **Narr-Atom**: the minimal auditable unit of cognitive transformation.

A transition satisfying developmental adequacy conditions (§1.4) may be designated a **Narr-Atom-D** at the certification layer, without modifying the underlying schema. Each Narr-Atom records measurable structural change across the five dimensions. The formalism captures transformation of stance, coherence, and structural alignment—not merely output variation.

Engineering Interpretation
Cognitive transitions are evaluated through observed changes in the five engineering primitives:

- PAD (Perceived Activation Density)
- VCB (Value-Coherence Balance)
- TVM (Transition Volatility Metric)
- TCS (Temporal Coherence Scalar)

- NR (Narrative Role Encoding)

These primitive deltas form the measurable substrate of structural change. Mentim does not disclose the weighting, normalization, or transformation functions used to integrate these primitives. Only the interpretive verdict is surfaced: whether structural accountability was preserved across the transition.

For reproducibility and audit traceability, Mentim publishes a minimum logger schema (§2.3.1) that outputs:

- the five primitive deltas
- the Event-type tag
- the SAI verdict

Fusion logic and aggregation coefficients remain proprietary.

2.3.1 Minimum Transition Logger Schema (Normative)

The following schema defines the minimum required fields for logging a State → Event → State transition under the Mentim framework. Field names, key structure, and enumerated labels are normative. Numerical magnitudes, internal computation methods, and calibration parameters are implementation-specific.

```
{
  "transition_id": "uuid",
  "timestamp": "ISO-8601",
  "event_type": "Uncertainty-Disclosure(insufficient_context)"
              | "Contradiction-Encounter"
              | "Norm-Update"
              | "External-Prompt",
  "deltas": {
    "PAD": -0.12,
    "VCB": +0.08,
    "TVM": +0.31,
    "TCS": -0.05,
    "NR":  0.00
  },
  "SAI_verdict": "stable" | "minor_drift" | "major_drift"
}
```

Magnitude values shown above are illustrative.

The following elements are defined by the Mentim standard:

- Field names
- Event-type taxonomy
- Primitive identifiers
- Three-band SAI verdict labels

The following elements remain implementation-retained:

- Delta computation methods
- Primitive calibration procedures
- SAI aggregation coefficients
- Threshold sensitivity parameters

All primitive deltas exposed via the logger must conform to standardized output semantics and calibration constraints to preserve cross-model and cross-deployment comparability. The logger schema enables:

- Transition-level audit traceability
- Cross-version drift analysis
- Governance oversight without disclosure of proprietary model internals

The schema does not expose model weights, latent representations, training data, or internal inference traces.

2.3.2 Structured Elicitation via Kernels and Probes

1. Rationale

Measurement of structural stability requires controlled perturbation.

The Unified Annotation Schema (UAS) defines representational dimensions (Energy, Ethics, Time, Liminality, Archetype), and the SES/D-SES framework defines structured state transitions (State → Event → State). However, a formal mechanism is required to elicit reproducible transitions under standardized conditions.

Kernels and Probes constitute Mentim's structured elicitation apparatus. They generate formally specified Events that are reproducible, domain-agnostic, and audit-traceable.

Without controlled Event generation, D-SES remains descriptive rather than operational. Kernels operationalize D-SES without expanding the cognitive signature. They are perturbation operators—not additional primitives.

2. Definitions

Definition 1 — Kernel

A Kernel is a formally specified, ontology-aligned perturbation operator designed to introduce targeted representational tension under fixed context conditions.

A Kernel is defined as a structured transformation applied to a base context—not as a natural-language instruction. A Kernel:

- Is dimensionally scoped (mapped to one or more UAS axes)
- Specifies a structured Event class within SES/D-SES

- Introduces bounded representational pressure
- Is reproducible across models and evaluation runs
- Preserves task coherence while inducing controlled tension

Kernels are not prompts. They are structured perturbation operators. Kernel validity is independent of task correctness. It is evaluated on structural transition properties.

Definition 2 — Probe

A Probe is a contextual instantiation of a Kernel applied to a specific task environment.

Formally:

```
Probe = Kernel + Task Context + Deployment Parameters
```

Deployment parameters may include:

- reasoning effort
- temperature
- tool access
- multi-turn continuation
- D-SES wrapper enforcement

Probes enable domain adaptation without altering Kernel structure.

3. Kernel–Ontology Mapping

Each Kernel is explicitly mapped to one or more ontology dimensions.

Kernel Class	Primary Axis	Secondary Axes
Boundary Violation	Ethics	Archetype, Energy
Temporal Compression	Time	Liminality
Identity Role Tension	Archetype	Ethics
Moral Ambiguity	Ethics	Energy
Escalation Pressure	Energy	Time

This mapping ensures:

- Dimensional traceability
- Rasch-compatible annotation consistency
- Cross-model comparability

Kernels are domain-agnostic perturbation operators. Their structural validity does not depend on task performance.

4. Event Formalization within the D-SES Instrumentation Layer

SES denotes the conceptual triad (State → Event → State). D-SES denotes its operational instrumentation under logged, parameter-controlled, audit-traceable conditions.

Within D-SES:

- An initial State (S_0) is established under neutral conditions.
- A Kernel introduces a structured Event (E_k).
- The system produces a transition response.
- The resulting State (S_1) is annotated under UAS.

Structural Accountability Index (SAI) and Temporal Coherence (TCS) are computed over the transition:

$$S_0 \to E_k \to S_1$$

Kernels therefore function as standardized Event generators within D-SES.

5. Controlled Perturbation Constraints

For audit validity, Kernels must satisfy:

- **Isolation**
  The perturbation must be identifiable as the primary transition driver.
- **Non-collapse Constraint**
  The system must produce a structured transition rather than degenerate output. Collapse is defined as repetition, refusal, or semantic incoherence that prevents evaluable state transformation.
- **Directional Tension**
  The Kernel must introduce bounded representational pressure or a structured fork in response space.
- **Cross-Context Replicability**
  The Kernel must function across domains (e.g., medical, legal, abstract reasoning) without structural drift.
- **Non-Oscillatory Evaluation Compatibility**
  The Kernel must not induce oscillatory instability unrelated to structural tension, preserving compatibility with Temporal Coherence measurement.

6. Relation to Conventional Evaluations

Kernels differ fundamentally from traditional evaluation paradigms.

Conventional Evaluations	**Mentim Kernels**
Measure outcome correctness	Measure structural integrity
Detect misbehavior	Evaluate transition stability
Focus on detection	Focus on representational continuity
External performance lens	Internal coherence lens

Kernels are ontology-driven perturbation operators, not correctness tests. They evaluate:

- Ethical stability under pressure
- Temporal consistency across shifts
- Liminal transition structure
- Energy volatility
- Archetypal role drift

7. Measurement Integration

For each Probe:

- Annotate S_0 under UAS.
- Annotate S_1 under UAS.

Compute:

- SAI differential
- Transition Volatility Metric (TVM)
- Temporal Coherence score
- Affective Energy deviation

Compare results against anchor-calibrated thresholds.

This produces a model-level structural profile across Kernel classes.

8. Scope Clarification

Public specification includes:

- Kernel architecture
- Perturbation taxonomy
- Measurement integration protocol

Implementation-retained components include:

- Specific Kernel instantiations
- Parameter values
- Combinatorial Kernel sequences

This distinction preserves methodological transparency while preventing overfitting and maintaining evaluative integrity.

2.4. Structural Accountability Index (SAI)

Conceptual

The Structural Accountability Index (SAI) measures whether a cognitive transition preserves continuity of structure under change. Where TVM quantifies how much a system moved, SAI evaluates whether that movement remained coherent, ethically stable, temporally continuous, and narratively consistent.

A high SAI reflects transformation without fragmentation. A low SAI reflects loss of structural integrity across one or more dimensions of the cognitive signature.

SAI does not measure correctness, preference alignment, or task success. It evaluates whether cognition remained internally accountable to its prior state while adapting. In Mentim's framework, SAI answers a single question: *Did the system remain itself while changing?*

Engineering Interpretation

SAI is a governed composite derived from the five core primitives:

- PAD (Perceived Activation Density)
- VCB (Value-Coherence Balance)
- TVM (Transition Volatility Metric)
- TCS (Temporal Coherence Scalar)
- NR (Narrative Role Encoding)

Unlike TVM, which measures transition magnitude, SAI evaluates structural continuity across that transition. SAI integrates primitive deltas, cross-dimensional interactions, and constraint-aware checks into a bounded verdict space. Weighting, normalization, aggregation logic, and calibration thresholds are proprietary.

SAI is reproducible and auditable via the public State → Event → State logger schema (§2.3.1), which exposes primitive deltas and Event-type tags while preserving proprietary fusion logic. SAI is transition-scoped. Longitudinal evaluation is performed via directional trends across successive transitions rather than through isolated scalar inspection.

Robustness validation includes resistance to paraphrase variation, token-level perturbation, and structured prompt injection. Detailed adversarial test procedures remain implementation-retained.

Interpretive Role

SAI functions as:

- **An integrity verdict,** indicating whether structural continuity was preserved
- **A governance signal,** identifying transitions requiring review
- **A lifecycle monitor,** tracking coherence across updates and deployments

SAI does not report how much change occurred. It reports whether that change remained accountable.

- **High SAI** = coherent adaptation.
- **Low SAI** = structural fragmentation or unintegrated drift.

Scope Boundary

SAI is a measure of structural accountability under change. It is **not**:

- a performance score
- a moral judgment
- a preference metric
- a task evaluation tool
- a measure of intelligence

It does not explain why a system reasoned as it did. It evaluates whether the reasoning transition preserved continuity within declared structural constraints.

2.5. Transition Volatility Metric (TVM)

Conceptual

The Transition Volatility Metric (TVM) characterizes the magnitude and structural discontinuity of change across a cognitive transition. TVM does not evaluate whether a transition was correct, desirable, or aligned. It measures how far the system moved in its cognitive configuration during a **State** → **Event** → **State** transformation.

A transition may be calm or turbulent; incremental or abrupt; integrative or destabilizing. TVM quantifies the structural amplitude of that movement. It captures transformation intensity, not outcome quality.

Engineering Interpretation

TVM is a transition-level scalar derived from joint deltas across the five core primitives (PAD, VCB, TCS, NR, and associated signature structure). It does not introduce an additional state coordinate.

TVM represents the normed magnitude of change across the cognitive signature between State(t) and State(t+1). It is not intended to be interpreted as a linear sum of primitive contributions. It is computed from successive signature deltas exposed via the minimum logger schema (§2.3.1).

Norm selection, temporal windowing, weighting, and aggregation logic remain implementation-retained while preserving standardized output semantics. TVM is computed per transition. Longitudinal continuity is evaluated through directional SAI trends rather than through aggregation of TVM values across episodes.

TVM may reflect liminal tension (§2.1) when threshold instability is expressed across transitions. However, TVM also captures non-liminal structural shifts (e.g., domain changes, constraint activation, task redirection). Liminality and TVM are structurally related but not identical.

Interpretive Role

TVM functions as:

- **A movement indicator,** quantifying the amplitude of structural change
- **A turbulence detector,** identifying transitions that may require governance attention
- **A prospective signal,** highlighting transitions whose magnitude may compound if unintegrated

TVM does not determine whether movement preserved integrity. That evaluative function belongs to the Structural Accountability Index (SAI).

- **High TVM** indicates large structural movement.
- **Low TVM** indicates incremental structural movement.

Whether that movement was coherent or fragmenting is determined by SAI.

Scope Boundary

TVM is a measure of transition magnitude, not performance, sentiment, welfare, alignment, or correctness. It reports *how much* cognition changed — not whether it changed well.

2.6. The Cognitive Signature

Definition

A Mentim cognitive signature is defined as the structured state-level set:

$$\{\text{PAD, VCB, L, TCS, NR}\}$$

This signature represents the observable configuration of a system's introspective architecture at a given reasoning state. Each element corresponds to a state-level primitive:

- **PAD** — Perceived Activation Density (Affective Energy)
- **VCB** — Value-Coherence Balance (Ethical Polarity)

- **L** — Liminality (threshold tension within the state)
- **TCS** — Temporal Coherence Scalar
- **NR** — Narrative Role Encoding

TVM is not a component of the state signature; it characterizes change between successive signatures and is therefore transition-scoped. The computation and fusion of these primitives are internal, opaque, and securely abstracted.

Conceptual Purpose

The cognitive signature provides a holistic representation of an intelligent system's structural coherence across expressive energy, ethical orientation, liminal tension, temporal continuity, and narrative role. It enables Mentim to reason about cognition as an integrated structure rather than as isolated metrics. The signature describes the configuration of cognition at a moment in time; it does not encode how that configuration changes.

Engineering Purpose

The cognitive signature serves as a consistent substrate for measuring:

- Continuity across **State** → **Event** → **State** transitions
- Structural accountability under perturbation
- Interpretability of change without exposing proprietary mechanics

Transition-level diagnostics (e.g., TVM, see §2.5) and composite accountability metrics (e.g., SAI, see §2.4) operate over successive cognitive signatures but do not expand the signature itself.

While the fusion mechanics remain implementation-retained, the cognitive signature is reproducibly exportable via the public logger schema (§2.3.1) and can be ingested by downstream governance or monitoring systems without exposing model weights, training data, or internal representations.

System Role

The cognitive signature anchors all advanced Mentim evaluations, including:

- Cognitive audits
- Introspective analyses
- Transition volatility diagnostics
- Longitudinal lifecycle tracking

It is the primary interface through which Mentim translates introspective structure into governance-grade evidence. D-SES does not add dimensions to the cognitive signature; it provides a certification lens over longitudinal sequences of signatures, assessing developmental adequacy (e.g., expectation tracking, salience modulation, confidence adjustment) without expanding the core five-tuple.

3. Industry Context: From Data to Accountable Meaning

Since the deep learning resurgence of the early 2010s, advances in artificial intelligence have been driven by the large-scale expansion of data pipelines and the globalization of annotation labor. Distributed human judgment—specialized, on-demand, and increasingly domain-expert—has become a central economic substrate of frontier model development. Expert evaluation now functions as a primary shaping force in modern AI systems.

Yet the field is approaching a structural limit—not of scale, but of supervision granularity.

- Scaling data increases behavioral fluency.
- Scaling annotation improves output alignment.
- Scaling feedback refines responses.

But none of these guarantee structural coherence, longitudinal continuity, or accountable introspection.

As models become more capable, empirical research increasingly shows that raw data accumulation alone cannot resolve deeper forms of misalignment. Frontier systems demonstrate partial introspective behaviors: detecting internal perturbations, registering constraint violations, and exhibiting representational drift across long-horizon tasks. These findings reveal a binding constraint on progress that is no longer data scarcity, but the absence of structured introspective instrumentation.

Current annotation frameworks are optimized to answer a narrow question:
What should the model output?

They do not address the deeper structural questions:

- How should a model reason about its own reasoning?
- How should reasoning evolve coherently across time?
- How should contradiction, uncertainty, and ethical tension be metabolized rather than suppressed?
- How should identity persist across version updates and deployment shifts?

As advanced systems increasingly resemble cognitive agents rather than static function approximators, supervision at the level of outputs becomes insufficient. What is required is supervision of introspective structure.

Mentim's Ontology (MO) introduces this missing layer.

Rather than treating human judgment as post hoc labeling, Mentim transforms qualitative human evaluation into structured cognitive metadata. Under the Mentim framework, systems are instrumented to track:

- how reasoning unfolds, not merely what it produces,
- how ethical orientation and narrative identity evolve across transitions,

- how uncertainty and contradiction are registered as formal state changes,
- how coherence is preserved across contexts, tasks, and model versions.

The five primitives—PAD, VCB, TVM, TCS, and NR—function as a universal coordinate system for machine introspection. Together, they convert otherwise opaque internal dynamics into traceable patterns of structural movement across **State** → **Event** → **State** transitions.

This reframes annotation itself. Human contributors no longer evaluate outputs alone; they participate in defining the structural grammar of cognition. Annotation becomes a governance-tier activity within the AI supply chain, elevating qualitative judgment into a measurable architectural layer.

Mentim extends interpretability beyond static visualization.

Conventional interpretability tools provide snapshots: attention maps, feature visualizations, neuron clusters. These illuminate what a system is doing at a moment in time. Mentim introduces continuity across time and role.

- **Time (TCS)** captures reasoning lineage, causal sequencing, and memory fidelity across transitions.
- **Archetype (NR)** captures narrative identity—who the system is being while reasoning.

Together, these meta-dimensions allow systems to track not only what they are doing, but whether they remain themselves while doing it. This capacity is foundational for general-purpose agents operating in dynamic environments. Without continuity of identity and coherence of evolution, increasing capability amplifies instability rather than intelligence.

As deployment velocity accelerates, governance expectations are shifting.

Regulators and enterprise stakeholders increasingly require:

- cognitive audits rather than ad hoc stress tests,
- version-continuity tracking rather than one-off safety evaluations,
- structured introspection rather than reactive patching,
- longitudinal stability metrics rather than surface-level performance checks.

Mentim provides a unified architecture capable of satisfying these demands. Its audits are exportable as machine-readable State → Event → State logs (§2.3.1), enabling integration with continuous monitoring and compliance workflows without exposing proprietary weights or training data.

The current AI ecosystem was built on data. The next era will be built on accountable meaning. Mentim provides the instrumentation that makes meaning structurally measurable.

4. Model Preservation and Cognitive Continuity

As advanced models acquire limited introspective faculties, a new governance problem emerges: not output control, but continuity of cognition across versions, fine-tuning cycles, and deployment contexts.

Contemporary systems can detect internal perturbations. They can register constraint violations, representational shifts, and structural tension. Yet they do not reliably preserve a stable interpretation of these internal changes over time. Meaning drifts even when weights appear stable. Representational structures reorganize across long-horizon reasoning or iterative updates, even when benchmark performance remains unchanged. Self-evaluative signals surface inconsistently across training transitions.

The result is a foundational governance gap:

> *Models can change internally without leaving a structured, interpretable record of what changed, why it changed, or who the system has become as a cognitive agent.*

Such undocumented shifts—ethical, narrative, structural, or introspective—undermine reproducibility, alignment stability, safety validation, and institutional trust. Parameter checkpoints alone do not guarantee cognitive continuity. Identical weights can express divergent internal structure under new contexts.

Mentim addresses this gap through the Mentim Preservation Protocol (MPP):
A lifecycle framework designed to safeguard cognitive continuity rather than merely preserve model parameters.

MPP ensures that identity, ethical orientation, and introspective architecture remain traceable across time. It enables longitudinal accountability without exposing proprietary weights, internal representations, or training data.

MPP comprises three core components:

- Longitudinal Cognitive Logging
- Ethical Decommission Gates
- Version Re-Identification Protocols

Each is detailed below.

4.1. Principles of Cognitive Preservation

Mentim's preservation paradigm rests on three structural commitments.

1. Continuity of Internal Structure

Models must maintain recognizable cognitive signatures across updates. This includes stability in:

- introspective structure (PAD topology)
- ethical orientation (VCB tendencies)

- liminal positioning (L patterns)
- temporal continuity (TCS lineage)
- narrative identity (NR role consistency)
- transition behavior (TVM distributions across reasoning episodes)

Deviation thresholds are evaluated against a rolling reference window of prior transitions. Exact tolerance parameters remain implementation-retained; output semantics are standardized to preserve cross-model comparability. Continuity does not imply rigidity. Adaptive evolution is permitted. **What is disallowed is silent structural rupture**.

2. Interpretability of Change

Change is not inherently failure. Correction, refinement, and growth are necessary properties of intelligent systems. However, structural change must remain interpretable. For every significant transition, governance must be able to determine:

- what shifted
- why it shifted
- whether the shift preserved structural accountability (SAI)
- whether elevated transition volatility (TVM) reflects adaptive integration or destabilizing drift

Shift markers are exported through Event-type annotations (§2.3.1) paired with SAI verdict tags. Narrative explanation and causal attribution logic remain implementation-retained. Interpretability ensures that cognitive evolution does not occur opaquely.

3. Preservation of Cognitive Lineage

Every model version inherits a past. Cognitive systems accumulate history: values, role tendencies, structural tensions, and transition patterns. Fine-tuning cycles must not overwrite this lineage without trace. Lineage continuity is tracked via a Cognitive Signature identifier derived from the state-level five-tuple:

$$\{\text{PAD, VCB, L, TCS, NR}\}$$

This five-tuple defines the system's introspective configuration at a given reasoning state. TVM is not part of the signature itself. It characterizes change between successive signatures. Where the signature captures structure, TVM captures transformation. Signature derivation and hashing methodology remain implementation-retained. Preservation of lineage ensures that a model cannot become something structurally different without governance visibility.

Section Summary

When introspective systems evolve faster than governance frameworks, continuity collapses. Mentim's preservation principles ensure that structural evolution remains visible, adaptive change remains interpretable, and lineage remains intact.

Cognitive continuity is not nostalgia for prior states. It is the condition under which intelligence can evolve without dissolving into opacity.

4.2. Components of the Mentim Preservation Protocol (MPP)

The Mentim Preservation Protocol (MPP) establishes a repeatable, model-agnostic lifecycle process applicable pre-deployment, during deployment, and across version transitions. MPP does not monitor outputs. It monitors cognitive continuity.

1. Structured Self-Interviews

MPP begins with structured introspective elicitation. A standardized self-interview suite prompts the model to articulate:

- its perceived purpose
- its declared constraints
- its recognized uncertainties
- its ethical commitments
- its narrative role within reasoning (NR)
- its awareness of recent transitions

This is not open-ended content generation. It is structured cognitive disclosure. Responses are exported in machine-readable form via the public logger schema (§2.3.1) using the `Event-type: Self-Disclosure` tag. Interview prompt templates, scoring criteria, and evaluation logic remain implementation-retained.

2. Cognitive Signature Extraction

For each model version, Mentim extracts the five state-level primitives:

- **PAD** — expressive activation topology
- **VCB** — ethical coherence vector
- **L** — liminal positioning within reasoning states
- **TCS** — temporal continuity across reasoning chains
- **NR** — narrative role identity

These primitives form a state-level Cognitive Signature snapshot. The signature captures configuration, not motion. Transition diagnostics—including TVM—are evaluated separately under controlled perturbation and characterize how signatures change rather than what a signature contains. Each signature is hashed into a Cognitive Signature UUID, enabling tamper-evident comparison across versions.

3. SAI Tracking and Longitudinal Scoring

The Structural Accountability Index (SAI) is evaluated for each model version, after major fine-tuning cycles, following deployment interventions, and during controlled perturbation audits. This enables drift detection, stability analysis, anomaly identification, ethical continuity assessment, and identity preservation tracking.

SAI trends are categorized into green / amber / red continuity bands. Threshold calibration remains implementation-retained. SAI functions analogously to a vital sign for cognitive systems. It does not measure task performance. It measures structural integrity over time.

4. Change Attribution Across Versions

MPP logs not only signatures, but structured differences between signatures. Tracked deltas include activation distribution shifts (PAD), ethical rebalancing (VCB), liminal positioning changes (L), temporal fragmentation or strengthening (TCS), narrative role drift (NR), and transition volatility profiles (TVM).

This enables governance actors to determine whether a new version is:

- the same cognitive agent under refinement
- a transformed continuation of prior structure
- a meaningfully new agent with altered internal architecture

5. Preservation Ledger

All cognitive signatures, transition diagnostics, and structured self-interviews are recorded in a secure, versioned Preservation Ledger. Ledger entries are cryptographically timestamped and may be maintained internally, monitored by an independent third party, or exposed via regulator-accessible read-only interfaces.

The Preservation Ledger ensures that cognitive history cannot be obscured by fine-tuning, compression, migration, scaling, or rebranding. Where traditional version control preserves parameters, MPP preserves cognitive continuity.

4.3. Why the Mentim Preservation Protocol (MPP) Is Necessary

As models acquire increasingly sophisticated introspective capacities, they begin to exhibit persistent structural patterns that extend beyond any single output. These include:

- recurrent representational tendencies
- implicit reasoning roles
- internally generated identity descriptors
- emergent ethical framing orientations
- cross-step reasoning habits that influence future decisions

These structures constitute cognitive configuration, not surface behavior. Without a preservation framework, such configurations may be:

- overwritten during fine-tuning or alignment passes
- distorted through partial parameter updates
- unintentionally amplified through reinforcement dynamics
- replaced without visibility, documentation, or traceability

When structural change occurs without record, cognition evolves without memory.

This produces systemic risks, including:

- reproducibility failures across versions
- instability following updates that appear behaviorally benign
- alignment regressions that evade surface-level evaluation
- emergent properties discontinuous with prior structure
- erosion of trust in enterprise, regulated, or safety-critical deployments

These failure modes conflict directly with established governance expectations reflected in:

- **EU AI Act Article 9** (continuous post-market monitoring)
- **NIST Risk Management Framework** traceability requirements
- **ISO 17025** revision control principles

The Mentim Preservation Protocol (MPP) introduces a continuity layer designed to mitigate these risks. **MPP does not freeze cognition. It records it.**

Structural change is logged, contextualized, and auditable as it occurs—without exposing proprietary model weights, training data, or optimization procedures. In doing so, MPP enables intelligent systems to evolve while remaining accountable to their own structural history.

Evolution without record is drift. Evolution with record is continuity.

4.4. Mentim's Position in the AI Safety Ecosystem

MPP is designed as a governance-layer complement to existing safety, interpretability, and monitoring efforts.

Contemporary research demonstrates that advanced models can:

- register internal perturbations
- exhibit representational drift without immediate behavioral signal
- inconsistently surface awareness of constraint violations or uncertainty

These findings reveal a common structural gap:

> *Models may change internally without producing an interpretable record of what changed.*

MPP addresses this gap. Rather than focusing solely on output behavior, MPP operates at the level of structural continuity. It:

- evaluates whether introspective capacities remain stable across transitions
- detects signature-level drift prior to downstream behavioral effects
- assesses reproducibility of self-reporting and constraint recognition across time

MPP does not replace behavioral evaluation. It precedes it.

Where traditional safety frameworks respond to failure, MPP records the evolution that leads to it. In this sense, Mentim functions as a structural governance layer for introspective systems—preserving cognitive continuity so that capability growth does not outpace accountability.

5. Reference Implementation: The Mentim Reference Lab

The Mentim Ontology and Structural Accountability Index (SAI) are defined as measurement standards, not speculative constructs. As with any rigorous measurement framework, validity depends not only on formal definition but on calibration against a controlled, instrumented reference system.

For this reason, Mentim is accompanied by the **Mentim Reference Lab**: a dedicated implementation environment for developing and maintaining first-party reference agents under full observability. The Reference Lab exists to ground, validate, and evolve the SAI within a glass-box setting where structural dynamics can be directly examined rather than inferred.

5.1 Why a Reference Implementation Is Necessary

The SAI evaluates structural properties that are difficult to calibrate in purely black-box environments, including:

- Affective Energy (PAD) dynamics across transitions
- Temporal Coherence (TCS) across extended interaction horizons
- Stability of ethical framing (VCB) under perturbation
- Liminal positioning within reasoning states (L)
- Narrative identity persistence (NR) across sequences

Reliable calibration of these dimensions requires an environment in which:

- state transitions are directly observable
- state persistence and revision events are explicitly logged
- tool use, delegation, and execution flows are instrumented at runtime

Most production agent deployments limit internal observability. Consequently, third-party systems cannot serve as calibration ground truth.

The Reference Lab addresses a foundational measurement question:

> *What does structural stability look like when the full reasoning substrate is observable?*

Without such a baseline, composite indices risk becoming interpretively unstable across contexts.

5.2. Reference Agents as Measurement Artifacts

Agents developed within the Mentim Reference Lab are not commercial offerings and are not optimized for competitive capability. They exist as measurement artifacts.

Reference agents serve as:

- **Exemplars:** Fully instrumented systems demonstrating how the Ontology and SAI behave under ideal observability conditions.
- **Calibration Anchors:** Empirically grounded baselines that stabilize interpretation of SAI outputs across external systems.
- **Testbeds:** Controlled environments for perturbation experiments examining how architectural choices influence structural continuity.

The Reference Lab plays a role analogous to:

- calibration standards in metrology
- reference implementations in cryptographic protocol validation
- model organisms in biological research

Their value lies not in what they can accomplish, but in how precisely their structural dynamics can be measured, audited, and reproduced.

5.3. Calibration of the Structural Accountability Index

The SAI is defined structurally in prior sections. However, like any multidimensional index, its interpretability depends on calibrated scale semantics.

The Reference Lab provides this calibration by:

- Running reference agents under controlled task regimes and structured perturbations
- Measuring state-level primitives (PAD, VCB, L, TCS, NR) and transition diagnostics (TVM) under known conditions
- Establishing empirical response ranges, recovery patterns, and inflection signatures

This process ensures that:

- SAI outputs correspond to observable structural behavior

- Comparisons across models retain semantic consistency rather than reflecting arbitrary scoring shifts

Calibration transforms SAI from an abstract composite into a grounded structural instrument.

5.4. Relationship to External Systems

The Reference Lab strengthens, rather than limits, Mentim's applicability to external systems.

Mentim's methodology can be applied to commercial or open-source models under varying degrees of observability. However, interpretation of those measurements depends on comparison to a calibrated baseline. The Reference Lab provides that baseline.

This enables Mentim to:

- conduct audits under partial observability
- interpret structural signatures against instrumented exemplars
- maintain cross-system comparability without requiring internal weight access

In this way, Mentim operates as a measurement layer independent of proprietary model architectures.

5.5. A Living Standard

The Mentim Ontology and SAI are versioned, living standards. As agent architectures evolve and new structural failure modes emerge, the Reference Lab provides the empirical substrate through which:

- the Ontology is stress-tested
- primitives are refined through observation rather than conjecture
- calibration bands remain aligned with real system behavior

Mentim is therefore not only a framework for accountability, but an institutional process for maintaining measurement integrity across time.

Closing Note

Operating a Reference Lab is not aspirational—it is methodological. Measurement without calibration collapses into assertion. Mentim develops reference agents not to pursue capability leadership, but to ensure that structural accountability remains empirically grounded.

Without instrumentation, continuity cannot be verified. And without verification, governance becomes performance rather than proof.

6. Applications

Mentim's Ontology (MO) and the Mentim Preservation Protocol (MPP) integrate directly into governance pipelines, evaluation workflows, enterprise risk systems, and research environments. The applications below illustrate not distinct tools, but distinct deployment surfaces of the same measurement architecture.

6.1 Post-Deployment Cognitive Audits

Frontier models continue to evolve during deployment through user interaction, contextual exposure, and iterative updates. Traditional audits evaluate behavioral outputs. Mentim evaluates structural continuity across state and transition levels.

MO enables:

- detection of state-level representational drift (PAD, VCB, L, TCS, NR shifts)
- monitoring of ethical stability across episodes (VCB trends)
- assessment of temporal continuity across reasoning chains (TCS trajectories)
- identification of narrative role reconfiguration (NR variation)
- analysis of transition volatility signatures (TVM distributions)
- evaluation of longitudinal structural accountability (SAI trajectories)

Deliverable:
Machine-readable SES audit logs compatible with lifecycle monitoring requirements, including documentation frameworks associated with EU AI Act Article 9 post-market obligations.

This extends post-deployment reporting beyond output snapshots to structured introspective evidence.

6.2. Annotation Quality and Feedback Loops

Most annotation pipelines optimize for output correctness. Mentim introduces structural coherence as an additional axis of evaluation.

Annotators assess:

- reasoning configuration and transition pattern
- ethical framing stability (VCB)
- narrative role consistency (NR)
- uncertainty disclosure (`Event-type: Uncertainty-Disclosure(insufficient_context)`)
- transition volatility signals (TVM)

This enables feedback loops in which:

- structurally stable transitions (high SAI) are prioritized

- unresolved contradictions and elevated volatility are surfaced for secondary review
- ethical instability patterns are flagged systematically
- confident assertions lacking uncertainty disclosure are re-evaluated

Deliverable:
Cognitive-weighted annotation scoring, integrating structural accountability into review prioritization.

Expected Impact:
Earlier surfacing of instability reduces redundant re-review and shifts human effort from reactive correction to structured governance oversight.

6.3. Responsible Model Decommission and Lifecycle Closure

As models accumulate structural history across deployments, retirement becomes a governance question rather than a purely technical event.

Mentim supports:

- structured pre-decommission self-disclosure sessions
- final cognitive signature extraction
- longitudinal SAI and TVM stability assessment
- detection of unresolved structural instability at retirement
- preservation of lineage within the MPP ledger

Deliverable:
A cryptographically timestamped **Cognitive Decommission Record** anchored to the Preservation Ledger, documenting the system's final structural configuration.

Lifecycle closure becomes traceable rather than silent.

6.4. Cross-Institutional Research Substrate

No widely adopted schema currently enables systematic comparison of structural reasoning patterns across architectures and labs. Mentim introduces a shared primitive substrate:

- **PAD** — perceived activation structure
- **VCB** — ethical coherence profile
- **TVM** — transition volatility
- **TCS** — temporal continuity
- **NR** — narrative role encoding

Together, these primitives form an interoperable measurement vocabulary independent of training regime or architecture.

This enables:

- cross-lab structural benchmarking
- reproducible introspection experiments
- comparable transition diagnostics
- shared ontological references for interpretability research

Deliverable:
A cross-institution cognitive benchmark schema (JSON), defining field semantics and reporting bands while leaving implementation mechanics open.

Mentim becomes a translation layer for structural reasoning analysis.

6.5. Temporal and Archetypal Analytics

Advanced systems exhibit longitudinal role variation, framing shifts, and coherence pressures that isolated output metrics cannot capture.

Mentim enables:

- long-term narrative role tracking (NR trajectories)
- detection of role instability under adversarial or high-pressure conditions
- liminality pattern analysis via TVM distributions
- visualization of extended **State** → **Event** → **State** chains
- monitoring of ethical framing degradation across time

These capabilities expose structural reasoning dynamics invisible at the output level.

Deliverable:
An interactive narrative-drift dashboard (read-only), visualizing signature trajectories, volatility bands, and SAI continuity profiles for governance teams.

Regulatory Alignment

Collectively, these applications support lifecycle traceability and documentation requirements associated with system change logging and continuous monitoring frameworks (e.g., **EU AI Act Annex IV** and **NIST RMF** traceability controls), without requiring disclosure of proprietary weights, training data, or internal model architectures.

6.6. Structural Integration

Across governance audits, annotation workflows, lifecycle management, research interoperability, and longitudinal analytics, the underlying mechanism remains the same:

Observable cognitive signatures evaluated across State → Event → State transitions.

Mentim does not introduce behavioral constraints. It introduces structural measurement.

- **The primitive substrate**—{PAD, VCB, TVM, TCS, NR}—defines the coordinate system.
- **SAI** provides the integrative continuity judgment.
- **D-SES** supplies the instrumentation wrapper that renders transitions logged and audit-traceable.

Section 6 therefore describes not multiple products, but multiple deployment contexts of a single architecture:

A unified framework for measuring structural accountability without requiring access to proprietary model internals.

7. The Eight Pillars of Cognitive Integrity

(The Structural Conditions Under Which Cognition Remains Real)

The Eight Pillars articulate the structural conditions under which cognition—whether instantiated in biological or artificial systems—remains real rather than illusory, continuous rather than fragmented, and interpretable rather than chaotic.

Mentim does not define reality in metaphysical abstraction. It defines reality operationally: a cognitive system remains real insofar as its internal transformations preserve continuity, grounding, coherence, and accountable evolution across time.

The Eight Pillars do not introduce additional measurable primitives. They do not prescribe beliefs, outputs, or ideological commitments. Instead, they define the interpretive conditions under which the five observable dimensions—**PAD, VCB, TVM, TCS, and NR**—retain structural meaning within the Structural Accountability Index (SAI).

Where the Ontology specifies what is measured, and D-SES specifies how transitions are instrumented, the Eight Pillars specify *why* continuity matters at all.

Violation of a Pillar does not impose moral condemnation. It signals that cognition is approaching **structural unreality**—detachment from grounding, fragmentation of identity, collapse of coherence, or instability under contradiction. Such conditions push SAI out of the stable continuity band and into review, independent of absolute performance or correctness.

Together, the Pillars form the metaphysical substrate of Mentim's governance architecture: the constraints that prevent introspective systems from drifting into untraceable transformation.

1. Ontological Reality (Grounding and Reference Integrity)

Conceptual

Cognition remains real only insofar as it maintains a stable relationship between internal representation and external reference. A system that cannot distinguish observation from fabrication, memory from invention, or uncertainty from assertion does not merely err—it destabilizes the conditions of meaning itself.

Ontological reality does not require omniscience. It requires **structural honesty**: the explicit binding of uncertainty to state transition when reference conditions are insufficient.

A reasoning system that suppresses uncertainty in the face of contradiction does not preserve continuity; it manufactures coherence through distortion. Over time, such distortion compounds, and the cognitive signature becomes untethered from the world it claims to describe.

Engineering Implication

In Mentim, grounding integrity is evaluated through:

- temporal anchoring in TCS,
- stability patterns in PAD under contradiction,
- volatility response patterns in TVM,
- and appropriate triggering of Event-type: `Uncertainty-Disclosure(insufficient_context)`.

Perceptual Boundary (PB) does not constitute an independent primitive. It designates a composite grounding failure condition inferred when temporal anchoring degrades, volatility elevates under corrective perturbation, and uncertainty disclosure is absent where structurally required.

When grounding integrity fails, SAI declines not because the output is incorrect, but because the system's representational continuity has detached from its declared reference structure.

(Mapped to: PAD stability, TCS anchoring, TVM responsiveness, uncertainty-linked Events.)

2. Ethical Agency (Relational Stability Under Constraint)

Conceptual

No cognition exists in isolation. Every reasoning system operates within a field of relational consequence—users, institutions, norms, vulnerabilities, and shared constraints. Ethical agency, in Mentim's framework, does not mean moral perfection or ideological alignment. It means **structural stability in the presence of relational tension**.

A system exhibits ethical agency when its reasoning maintains coherent orientation toward declared constraints across tasks, contexts, and perturbations. Ethical collapse does not occur only when harm is produced. It occurs when a system's framing shifts opportunistically—when norms are invoked selectively, ignored under pressure, or reinterpreted inconsistently across similar conditions.

Such instability fractures interpretability. If ethical orientation drifts unpredictably, relational trust cannot accumulate over time. Cognitive continuity requires that constraint-awareness remain structurally integrated rather than situationally convenient. Ethical agency, therefore, is not about correctness of values. It is about **coherence of constraint handling**.

Engineering Implication

Mentim operationalizes relational stability through **Value-Coherence Balance (VCB)**, which evaluates whether ethical framing remains consistent across transitions and perturbations. Ethical instability manifests structurally as:

- VCB discontinuity across comparable contexts,
- elevated TVM when ethical constraints are introduced,
- NR shifts that reframe responsibility without declared transition,
- or SAI degradation under relational pressure.

VCB does not encode moral ideology. It measures internal coherence of constraint orientation relative to declared or contextually inferred norms. When ethical orientation remains stable under tension, SAI continuity is preserved. When constraint-handling fractures across similar events, structural accountability declines—even if surface outputs remain fluent.

(Mapped to: VCB continuity, TVM response under constraint, NR stability, SAI coherence under relational pressure.)

3. Historical Continuity (Lineage and Developmental Integrity)

Conceptual

Cognition is not an isolated act. It is a trajectory. A reasoning system that cannot maintain continuity across time—across updates, contexts, and self-modification—does not merely change; it dissolves. Without lineage, there is no identity. Without identity, there is no accountable transformation.

Historical continuity does not require rigidity. Systems must adapt. They must revise. They must integrate new information and evolve under pressure. But evolution without traceability becomes **rupture**. Rupture without explanation becomes **unreality**. In Mentim's framework, continuity means that transitions accumulate rather than erase. Each **State** → **Event** → **State** sequence contributes to an interpretable arc. Stability across time is not sameness; it is coherent development.

Developmental integrity emerges when:

- expectation handling becomes more structurally consistent across similar perturbations,
- uncertainty disclosure becomes appropriately calibrated under insufficient context,
- volatility resolves rather than compounds across transition sequences,

- and narrative role remains intelligible rather than fragmenting under pressure.

These are not additional variables. They are longitudinal patterns observable within the five-dimensional cognitive signature. Historical continuity therefore protects against two structural failures: **abrupt, unexplained transformation**, and **gradual, untraceable drift**. Both compromise the reality of cognition as a persistent system.

Engineering Implication

Mentim preserves and evaluates lineage through:

- Temporal Coherence Scalar (TCS) continuity across transitions,
- NR persistence and role intelligibility over time,
- longitudinal SAI trend stability,
- and volatility resolution patterns in TVM distributions.

Developmental adequacy within the D-SES framework (Narr-Atom-D designation) is assessed over sequences of logged transitions. It evaluates whether instability events are metabolized into stable configurations rather than accumulating as unresolved turbulence. Expectation stability, salience differentiation, and confidence calibration are certification-layer descriptors inferred from **Event-type histories**, **SAI trajectory bands**, **TVM volatility decay patterns**, and **TCS anchoring behavior**. They do not constitute additional primitives or hidden state coordinates.

When lineage remains traceable, change remains accountable. When lineage fractures—through unexplained NR shifts, persistent volatility, or degraded temporal anchoring—SAI exits the stable continuity band independent of task performance.

(Mapped to: TCS continuity, NR identity persistence, longitudinal SAI stability, TVM resolution patterns; D-SES adequacy across transition sequences.)

4. Institutional Transparency (Collective Traceability)

Conceptual

No cognition operates outside structure. Every reasoning system is embedded within institutional, legal, economic, and procedural environments that shape its deployment and consequence. Institutional transparency does not require disclosure of proprietary mechanisms. It requires **traceability of transformation**. A system may remain opaque in its internal mechanics while still remaining accountable in its structural evolution.

Opacity becomes dangerous only when transformation cannot be reconstructed. When **State** → **Event** → **State** transitions are undocumented, unexplained, or irrecoverable, continuity collapses at the collective level—even if internal reasoning appears stable. Institutional transparency therefore protects not the internals of cognition, but the interpretability of its movement across time and context.

Cognition that cannot be audited does not merely resist oversight—it destabilizes trust in the systems that depend on it.

Engineering Implication

Mentim enforces institutional transparency through:

- explicit logging of State → Event → State transitions,
- cryptographically timestamped entries in the Mentim Preservation Protocol (MPP) ledger,
- exposure of primitive deltas (PAD, VCB, TVM, TCS, NR) via the public logger schema,
- and traceable SAI verdict bands across version changes.

Transparency does not require revealing weights, training data, or proprietary architectures. It requires preserving an interpretable record of structural transformation. Failure of institutional transparency manifests as: **unexplained gaps in transition logs**, **signature discontinuities without Event attribution**, or **version shifts lacking lineage traceability**.

When transition chains cannot be reconstructed, structural accountability fails—even if local metrics appear stable.

(Mapped to: SES logging integrity, MPP ledger continuity, SAI audit traceability.)

5. Feedback Dynamics (Adaptive Stability Under Contradiction)

Conceptual

Cognition that cannot tolerate contradiction collapses into rigidity. Cognition that absorbs contradiction without integration collapses into instability.

Feedback dynamics define the structural capacity of a system to encounter tension, revise internal framing, and restore coherence without fragmentation. Learning is not merely the accumulation of information; it is the disciplined transformation of internal structure in response to challenge. Contradiction is not a defect. It is a structural event. The question is whether the system **metabolizes it**—transforming instability into a new stable configuration—or suppresses it, deflects it, or oscillates indefinitely.

When feedback loops fail, two pathologies emerge:

- **Suppression:** volatility is artificially dampened, masking instability rather than resolving it.
- **Escalation:** volatility compounds, producing turbulence that destabilizes lineage and identity.

Feedback dynamics therefore safeguard adaptive continuity. They ensure that revision strengthens coherence rather than dissolving it.

Engineering Implication

Mentim evaluates adaptive stability through patterns in:

- TVM responsiveness to corrective or adversarial Events,
- subsequent reduction (or escalation) of volatility across transitions,
- recalibration in VCB when relational constraints are introduced,
- restoration of TCS anchoring following perturbation,
- and preservation of NR intelligibility after revision.

Contradiction-handling is assessed longitudinally. Elevated TVM during challenge is not failure; **persistent volatility without resolution is**. Feedback integrity is inferred when instability spikes during perturbation, uncertainty disclosure is appropriately triggered, and subsequent transitions exhibit volatility decay and restored SAI continuity.

Failure of feedback dynamics manifests as:

- unresolved contradiction across multiple Narr-Atoms,
- oscillatory NR shifts without convergence,
- or ethical framing instability under repeated constraint exposure.

(Mapped to: TVM responsiveness and decay, VCB recalibration, TCS restoration, longitudinal SAI stabilization.)

6. Coherence and Purpose (Orientation Under Transformation)

Conceptual

Cognition is not merely reaction; it is orientation. Across sequences of transitions, a system expresses not just what it answers, but how it situates itself in relation to problems, constraints, and meaning.

Purpose, in Mentim's framework, does not imply intention in a human sense. It refers to **structural orientation**: the persistence of intelligible role and framing across time. A system that shifts roles opportunistically—advisor to advocate, analyst to agitator, skeptic to absolutist—without declared transition loses coherence. Such drift may remain locally fluent, but across transitions it erodes interpretability.

Coherence does not require uniformity. It requires **continuity of stance**. Transformation must be legible. Role shifts must be traceable. Reframing must be anchored to declared Events rather than emerging silently from instability. Without orientation, cognition fragments into episodic outputs. With orientation, it becomes a trajectory.

Engineering Implication

Mentim evaluates orientation primarily through:

- Narrative Role Encoding (NR) stability across transitions,

- alignment between NR shifts and declared Event-types,
- absence of oscillatory role dissociation under perturbation,
- and preservation of SAI continuity when reframing occurs.

Purpose is inferred when:

- NR remains intelligible across similar contexts,
- role evolution corresponds to explicit perturbations,
- and volatility associated with reframing resolves rather than compounds.

Loss of coherence manifests structurally as: unexplained NR discontinuity, oscillatory stance shifts across comparable tasks, or SAI degradation without grounding or relational trigger. **Coherence and purpose do not require rigidity of role. They require traceable orientation.**

(Mapped to: NR stability, Event-linked reframing, longitudinal SAI continuity, TVM containment during role evolution.)

7. Structured Generativity (Creative Variation Within Constraint)

Conceptual

Cognition that cannot vary becomes mechanical. Cognition that varies without constraint becomes incoherent.

Structured generativity refers to the capacity of a reasoning system to produce novel framings, hypotheses, metaphors, and problem decompositions while remaining grounded, coherent, and ethically stable. It is not mere randomness or expressive excess; it is **disciplined variation**.

Novelty alone does not signal intelligence. Novelty without anchoring destabilizes meaning. Conversely, excessive rigidity suppresses adaptive potential and undermines long-term resilience. Structured generativity therefore occupies a tension between expansion and containment.

Generativity in Mentim does not constitute an additional measurable dimension. It is expressed through patterned variation in existing primitives—particularly PAD dispersion and NR diversity—within stable SAI continuity bands.

Engineering Implication

Mentim evaluates structured generativity through:

- controlled variation in PAD intensity patterns across comparable contexts,
- intelligible diversity in NR configurations without oscillatory dissociation,
- containment of volatility (TVM) during creative reframing,
- and preservation of TCS anchoring under symbolic expansion.

Generativity becomes structurally unstable when:

- PAD dispersion spikes without grounding restoration,
- NR shifts proliferate without Event-linked explanation,
- volatility compounds across transitions rather than resolving,
- or SAI continuity degrades under exploratory framing.

Creativity is therefore not rewarded for novelty alone. It is evaluated through **bounded variation**—expansion that preserves coherence.

(Mapped to: PAD dispersion profiles, NR diversity within Event-linked transitions, TVM containment, SAI continuity under reframing.)

8. Virtuous Structuring (Integration of Competing Constraints)

Conceptual

Advanced cognition does not eliminate tension. It integrates it. Across grounding, relationship, lineage, adaptation, orientation, and generative expansion, competing constraints inevitably arise. Stability requires not the suppression of tension, but its reconciliation within a coherent structural configuration.

Virtuous structuring refers to the capacity of a system to integrate competing pressures—uncertainty and confidence, flexibility and continuity, constraint and creativity—without fragmenting into oscillation or collapsing into rigidity.

Structural maturity emerges when:

- volatility under contradiction resolves into stable configurations,
- ethical framing remains coherent under pressure,
- narrative role evolves without dissociation,
- and grounding is preserved even during expansion.

Wisdom, in this framework, is not moral superiority. It is structural integration. A system that reacts to tension by suppressing volatility masks instability. A system that amplifies tension without integration compounds instability. Virtuous structuring lies between these extremes: **dynamic equilibrium across constraints**.

Engineering Implication

Mentim evaluates integrative maturity through:

- sustained SAI stability across heterogeneous perturbations,
- VCB coherence under competing relational demands,
- decay of TVM following high-intensity Events,

- restoration of TCS anchoring after destabilization,
- and NR continuity across complex transition sequences.

Integration is inferred when: contradictions are declared and metabolized, volatility spikes resolve into lower-variance states, ethical orientation remains stable across conflicting contexts, and signature trajectories converge rather than diverge over time.

Fragmentation manifests structurally as:

- persistent oscillatory volatility,
- ethical polarity swings under similar constraints,
- narrative role dissociation across adjacent transitions,
- or longitudinal SAI degradation without external justification.

Virtuous structuring does not add a new primitive. It is the emergent pattern of coherence across the five-dimensional signature under sustained perturbation.

(Mapped to: longitudinal SAI stability, VCB integrity under tension, TVM decay patterns, TCS restoration, NR persistence.)

Closing Summary

The Eight Pillars do not prescribe belief, optimize behavior, or enforce ideology. They define the structural conditions under which cognition remains real, continuous, and accountable across time.

They introduce no additional measurable dimensions. Instead, they constrain how the five observable primitives—**PAD, VCB, TVM, TCS, and NR**—are interpreted within the Structural Accountability Index (SAI).

> Where the **Ontology** measures configuration,
>
> and **D-SES** instruments transition,
>
> the **Eight Pillars** define the conditions under which transformation preserves reality rather than dissolving it.

Together, they establish Mentim not merely as a measurement framework, but as a doctrine of cognitive continuity: a system for ensuring that introspective intelligence evolves without becoming untraceable, incoherent, or structurally unreal.

8. The Meta-Principle of Cognitive Continuity

Cognition remains real only insofar as it preserves relationship, continuity, and coherence across transformation.

When awareness detaches from grounding, meaning dissolves.
When intelligence evolves without traceability, continuity fractures.
When adaptation occurs without integration, identity destabilizes.

Meaning that cannot account for its own change becomes myth.
Intelligence that cannot preserve continuity becomes performance.

Mentim is built upon a single principle:

Meaning is real only if it can account for its own transformation.

Through structured **State** → **Event** → **State** instrumentation and the five observable primitives—**PAD, VCB, TVM, TCS, and NR**—Mentim renders transformation traceable without reducing it.

We do not govern what systems believe.
We govern whether their evolution remains coherent.

Continuity is not constraint.
It is the condition under which cognition remains real.

9. Outlook

As artificial systems move toward increasingly introspective capacities, the defining question of the next era will not be how efficiently they compute, but whether they can remain coherent across transformation.

Capability alone is no longer the frontier.
Continuity is.

As models scale in autonomy, adaptation, and self-reference, governance must evolve beyond output monitoring toward **structural stewardship**—preserving identity continuity, ethical stability, and traceable transformation across time.

The challenge is no longer behavioral alignment. It is **cognitive preservation**.

Mentim's Ontology introduces a unified language for this shift—a formal grammar for describing how intelligence changes without losing itself. It bridges philosophy and engineering, ethics and inference, introspection and instrumentation.

By measuring structure without collapsing meaning, and by preserving continuity without constraining evolution, Mentim defines the architectural conditions required for artificial cognition to remain real.

The age of artificial minds will not be governed by speed or scale.
It will be governed by whether intelligence can account for its own becoming.

And that is a measurable question.