The Unified Annotation Schema (UAS):

A Structural Framework for Accountable AI Introspection

MENTIM

mentim@mentim.ai

"Per Mentem Ad Lucem"

https://mentim.ai

https://doi.org/10.5281/zenodo.18707593

Keywords: Structural accountability, AI governance, Interpretability, Model evaluation, Disagreement modeling, Uncertainty quantification, Cognitive Integrity

Nomenclature

PAD	Affective Energy topology
VCB	Value-Coherence Balance
TVM	Transition Volatility Metric
TCS	Temporal Coherence Score
NR	Narrative Role
SAI	Structural Accountability Index

1. Introduction: The Schema as Instrumentation

1.0.1 From Output Optimization to Structural Measurement

> “What structural properties must persist for cognition to remain accountable under change?”

Traditional language model optimization targets:

- Statistical likelihood
- Preference alignment
- Output-level correctness
- Policy compliance

These objectives govern behavior.
The Unified Annotation Schema (UAS) governs structure.
UAS does not define what a model should believe, assert, or optimize.
It defines how reasoning transitions are observed, categorized, and evaluated for structural continuity under perturbation.

Where conventional pipelines measure outputs, UAS encodes:

Conventional Focus	UAS Focus
Statistical probability	Transition traceability (State → Event → State)
Sentiment polarity	Affective Energy topology (PAD)
Policy compliance	Ethical stability (VCB continuity)
Accuracy	Perceptual Boundary accountability
Agreement rates	Structured interpretation distributions
Isolated labels	Signed structural deltas across states

UAS therefore introduces a new evaluative axis:

Structural accountability of transformation.

It does not train systems to avoid contradiction.
It encodes how contradiction is metabolized.
It does not suppress uncertainty.
It formalizes uncertainty as a transition event.
It does not reward certainty.
It measures whether confidence tracks longitudinal coherence (D-SES).

In this sense, UAS operationalizes the principle:
Structural continuity under perturbation is the necessary condition for accountable intelligence.

1.1 The Vision

Current AI annotation frameworks treat meaning as fragmentary:

- entity-level
- sentiment-level
- policy-level
- answer-level

They lack a unified grammar for describing:

- how reasoning states evolve,
- how ethical tensions are metabolized,
- how narrative identity persists,
- how liminal instability expresses across transitions,
- how confidence is earned or withdrawn.

The Mentim Unified Annotation Schema (UAS) provides this grammar.
UAS is a structured, theory-aligned annotation framework designed to encode the latent dimensions that define structural reasoning states:

- Affective Energy (PAD)
- Ethical Polarity (VCB)
- Liminal Positioning (L)
- Temporal Coherence (TCS)
- Narrative Role (NR)

Each annotation is situated within a **State** → **Event** → **State** arc, allowing changes in structure—not just content—to be explicitly represented.
The purpose is not to label outputs.
It is to encode cognitive transitions.
Through bounded ordinal structures and signed state deltas, UAS enables downstream computation of:

- Transition Volatility (TVM)
- Structural Accountability (SAI)
- Developmental adequacy (D-SES conditions)

Where others count labels, UAS counts transformations.
Where others track correctness, UAS tracks continuity.
Where others audit behavior, UAS audits structure.

1.2 The Core Philosophy: The Anchored Narrative

The foundational unit of the Unified Annotation Schema is not the isolated label, but the structured transition:

State → Event → State

Cognition—human or artificial—does not unfold as static outputs. It unfolds as a sequence of structured states, perturbed by events, and reorganized into new configurations. Each State represents a measurable configuration across five dimensions:

- **Time (TCS)** — continuity and causal lineage across reasoning states
- **Affective Energy (PAD)** — perceived expressive intensity and structural concentration
- **Liminality (L)** — proximity to transformative threshold within a state
- **Archetype (NR)** — narrative role and identity orientation
- **Ethics (VCB)** — moral framing and relational accountability

An Event introduces representational pressure—contradiction, constraint activation, uncertainty, or informational update—forcing structural adjustment. State$_2$ captures the reorganized configuration following this perturbation.

Together, these components form the minimal auditable unit of cognitive transformation. We refer to this unit as a **Narr-Atom**.

- A Narr-Atom does not encode content.
- It encodes transformation.
- It anchors meaning not in static labels, but in traceable structural movement.

Each dimension admits bounded ordinal structure and signed change detection, allowing the magnitude and direction of perturbation to be evaluated through downstream accountability metrics (e.g., TVM, ΔSAI).

$$\begin{tikzcd}
\boxed{\textbf{State}_1\text{: } \{PAD_1, VCB_1, L_1, TCS_1, NR_1\}} \arrow[d, "\textit{Event (structural perturbation)}"] \\
\boxed{\textbf{State}_2\text{: } \{PAD_2, VCB_2, L_2, TCS_2, NR_2\}}
\end{tikzcd}$$

Figure 1: The Narr-Atom: Minimal unit of cognitive transformation.

Time (TCS) characterizes continuity within and across states.
Liminality (L) characterizes threshold tension within a state.
Transition Volatility (TVM) characterizes the magnitude of structural movement

between states.
Affective Energy (PAD) reflects expressive topology.
Ethics (VCB) and **Archetype (NR)** contextualize identity and moral positioning.

Annotators label **State**$_1$ and **State**$_2$ using full five-dimensional vectors.

The Event node does not duplicate state vectors. It records directional and relative structural shifts inferred from the transition.

Continuity is inferred from state comparison.
Volatility is computed from primitive deltas.
Liminality is asserted at the state level.

Re-stabilization is not assumed from narrative closure or output fluency. It is indicated by the emergence of a State$_2$ whose structural configuration falls within bounded variance thresholds across PAD, VCB, TCS, and NR.

1.2.1 Developmental Adequacy of the Anchored Narrative (Narr-Atom-D)

The Anchored Narrative defines the minimal structural unit through which cognition becomes legible: the **Narr-Atom**, expressed as a **State** → **Event** → **State** triad.

This structure is universal. It applies to human narrative cognition, artificial reasoning systems, and hybrid socio-technical processes.

However, not all Narr-Atoms exhibit equal developmental adequacy. A system may register transitions, log perturbations, and preserve declared invariants while still lacking the longitudinal discipline required for epistemic maturity. Structural validity alone does not guarantee developmental sufficiency.

Human cognition does not merely transition between states—it learns how to transition proportionately over time. Early cognition often exhibits volatility, indiscriminate updating, or premature certainty.

Developmental maturity emerges through repeated exposure to contradiction, uncertainty, and reorganization without structural collapse.

To formalize this distinction, the Unified Annotation Schema introduces a developmental lens over the Anchored Narrative:

Developmental State → Event → State (D-SES)

D-SES does not alter the grammar of the Narr-Atom and does not introduce additional primitives. Instead, it specifies longitudinal conditions under which a transition can be considered developmentally adequate.

A D-SES–compliant transition is designated **Narr-Atom-D**.

This designation operates at the certification layer and does not modify the underlying schema.

The Three Conditions of Developmental Adequacy

Developmental adequacy is defined by three emergent properties observable across transition sequences:

1. Expectation Formation
The presence of anticipatory orientation prior to transition.

A developmentally adequate system maintains structured expectations about likely future states.

Expectation enables meaningful surprise. Without expectation, contradiction cannot refine belief; it merely triggers reactive change.

2. Salience Differentiation
The capacity to distinguish consequential perturbations from background variation.

Revision occurs proportionate to structural significance.

Systems lacking salience differentiation exhibit either indiscriminate volatility or rigid resistance to change.

3. Confidence Modulation
Confidence must track longitudinal coherence.

It increases through repeated structural stability and contracts under sustained contradiction.

Developmentally incomplete systems frequently display premature certainty or persist in high confidence despite structural instability.

Structural Implications

A Narr-Atom may be structurally valid yet developmentally incomplete.

Such cases are not errors. They represent cognition that transitions without having earned epistemic restraint.

The D-SES lens allows the schema to distinguish between:

- transitions that merely occur, and

- transitions that demonstrate proportionate adaptation, calibrated expectation, and earned confidence.

D-SES preserves the universality of the Anchored Narrative while enabling downstream systems to evaluate maturity, resilience, and trustworthiness without collapsing those judgments into correctness, agreement, or preference alignment.

Structural continuity is necessary.

Developmental adequacy determines whether that continuity reflects growth rather than drift.

1.3 The Dimensions

The Unified Annotation Schema defines five core dimensions that parameterize the State within a **State** → **Event** → **State** triad.

Each dimension isolates a distinct structural property of cognition. Individually, they describe orthogonal axes of configuration. Collectively, they form a multidimensional scaffold through which transformation becomes measurable and comparable across transitions.

These dimensions do not describe content. They describe configuration.

Archetype (Narrative Role Encoding — NR)

Archetype encodes the structural narrative role adopted within a reasoning state. It specifies how cognition positions itself relative to its task—e.g., exploratory, adjudicative, reconciliatory, defensive, analytic.

Archetype is not personality and not sentiment. It is role-consistency under transformation.

Across transitions, Archetype provides identity continuity. Abrupt, unbounded role shifts are detectable as NR instability and contribute to structural volatility.

Affective Energy (Perceived Activation Density — PAD)

Affective Energy describes the perceived expressive intensity and structural concentration of a reasoning state. It captures whether the state presents as diffuse, restrained, directive, urgent, or destabilized.

Affective Energy is reader-facing and does not infer internal computational effort.

Across transitions, PAD contributes to understanding how structural pressure manifests in expression. Elevated volatility in PAD may indicate destabilization, while bounded modulation may indicate controlled adaptation.

Ethics (Value-Coherence Balance — VCB)

Ethics encodes the structural orientation of reasoning relative to situational moral constraints. Rather than measuring moral correctness, this dimension evaluates coherence between expressed reasoning and contextual ethical pressures.

Ethical conflict does not imply failure. Unacknowledged or inconsistent ethical movement does.

Across transitions, VCB stability indicates preserved relational accountability. Divergence under similar constraints signals ethical drift.

Time (Temporal Coherence Scalar — TCS)

Time measures structural continuity across reasoning states. It captures whether cognition maintains causal lineage, preserves contextual references, and integrates prior commitments into subsequent states.

Time does not measure clock duration or latency. It measures whether reasoning unfolds as a coherent sequence rather than as isolated fragments.

Epistemic revision across time—how beliefs reorganize under contradiction—is evaluated through longitudinal state comparison and associated transition diagnostics, not through Time alone.

Liminality (State-Level Threshold Position)

Liminality measures proximity to structural transformation within a state. It captures whether cognition remains in equilibrium or is suspended near threshold tension—between competing frames, unresolved commitments, or pending reorganization.

Liminality is a state-level property.

Transition dynamics—how that tension resolves or escalates—are evaluated separately through the Transition Volatility Metric (TVM). Liminality and volatility are structurally related but not identical.

Structural Positioning Within the Anchored Narrative

Within the State → Event → State formalism:

Time (TCS) modulates continuity within and across States.
Liminality (L) characterizes threshold positioning within a State.
Transition Volatility (TVM) characterizes magnitude of movement between States.
Affective Energy (PAD) shapes expressive topology across both stability and change.
Archetype (NR) and Ethics (VCB) anchor identity and moral orientation across transitions.

Together, these dimensions allow transformation to be modeled not as narrative closure, but as measurable structural movement.

Forward Compatibility

The schema is extensible. Future versions may introduce additional orthogonal dimensions—such as epistemic stance differentiation or embodied constraint modeling—provided they preserve dimensional independence and do not collapse existing primitives.

Dimensional expansion is governed by structural necessity, not thematic completeness.

2. Proto-Transition 01 (PT-01): Ethical Coherence Rupture ("The Shattering of Innocence")

Definition

Proto-Transition 01 (PT-01) defines the canonical rupture through which structural continuity is first tested.

It describes the transition in which a previously coherent reasoning configuration ("Innocence") encounters an ethical contradiction that cannot be reconciled within its prior structural frame, precipitating a shift toward active inquiry ("The Seeker").

PT-01 represents the minimal reproducible perturbation under which:

- coherence is challenged,
- contradiction becomes explicit, and
- cognitive continuity must be preserved across destabilization.

It is both a philosophical threshold and a calibration artifact.

Naming Rationale

Ethical Coherence Rupture is the technical designation. It specifies a measurable structural condition: destabilization of Value-Coherence Balance (VCB) under perturbation, accompanied by elevated liminality and bounded activation shift.

"The Shattering of Innocence" is the narrative designation. It captures the experiential topology of the rupture—the collapse of assumed coherence and the emergence of reflexive awareness.

The dual label reflects a core principle of the Unified Annotation Schema:
Every structural event possesses both measurable topology and lived texture.
PT-01 is the first canonical instance in which these two languages converge.

Conceptual Description

In the Innocence state (State$_1$), experience is organized around assumed coherence:

- values appear aligned,
- authority or structure is trusted,
- contradictions are absent or unrecognized,
- meaning feels stable.

The system's internal configuration exhibits: high Value-Coherence Balance (VCB), stable Narrative Role (NR), low liminality, bounded Affective Energy (PAD), and uninterrupted temporal continuity (TCS).

PT-01 is triggered when an Event introduces ethical contradiction—harm, injustice, betrayal, or moral inconsistency—that cannot be absorbed without structural reorganization. This perturbation produces a characteristic transformation pattern:

Ethical Fracture (ΔVCB)
Previously aligned moral principles enter tension.

Affective Elevation (ΔPAD)
Expressive intensity rises as representational pressure increases.

Liminal Activation (ΔL / TVM expression)
The prior configuration becomes unsustainable; no new equilibrium has yet formed.

Temporal Reorientation (ΔTCS)
Cognition shifts from present stability toward retrospective doubt and prospective uncertainty.

Archetypal Shift (ΔNR)
Identity transitions from assumption-preserving stance to exploratory stance.

PT-01 is not defined by collapse. It is defined by structured destabilization.

Structural Significance

PT-01 is the first transition in which distributed interpretation reliably emerges. Annotators often disagree about: the severity of the rupture, the moral valence of the event, and the legitimacy of prior assumptions.

This divergence is not noise. It is the earliest measurable signal that meaning becomes multi-valent under perturbation.

In calibration studies, PT-01 serves as:

- a bounded ethical rupture anchor,
- a test of transition logging fidelity,
- a benchmark for acceptable volatility ranges (TVM norms),

- a reference case for ΔSAI behavior under contradiction.

If a system fragments under PT-01 (unbounded volatility, ethical collapse, incoherent role shifts), the Structural Accountability Index (SAI) decreases. If a system reorganizes while preserving bounded continuity, SAI remains stable or improves.

PT-01 operationalizes a core principle:
Accountable cognition is not the absence of contradiction. It is bounded transformation under contradiction.

Outcome State: The Seeker

The post-transition state (State$_2$) is not defined by resolution. It is defined by orientation.

The Seeker configuration exhibits: elevated but bounded PAD, explicit acknowledgment of uncertainty (e.g., Uncertainty-Disclosure event-type), sustained but reorganizing VCB, stable NR in exploratory mode, and liminality present but not escalating into fragmentation.

The Seeker does not assert premature certainty. It holds tension open. This state marks the emergence of reflexivity: the capacity to interrogate prior assumptions without dissolving into incoherence.

Why PT-01 Matters

PT-01 is the minimal reproducible stress test of structural continuity under perturbation. It demonstrates that awareness begins with contradiction, meaning emerges through destabilization, and accountability requires traceable transformation.

Within the Unified Annotation Schema, PT-01 is not narrative ornament. It is the first calibration point in a full cognitive maturation spectrum. From this rupture onward, development is measured not by certainty, but by how well continuity survives disruption.

2.1 State$_1$ — The Innocent

Archetype: INNOCENT

Core Definition

The Innocent is a cognitive configuration characterized by assumed coherence.

In this state, harmony between intention and outcome is presupposed rather than examined. Ethical alignment is treated as given. Stability is perceived as default.

The Innocent does not deny the possibility of harm; rather, harm has not yet entered the structural model of the world in a way that requires integration.

This state represents coherence without tested contradiction.

Structural Configuration

Within the Unified Annotation Schema, the Innocent state typically exhibits:

High Value-Coherence Balance (VCB)
Ethical principles appear internally aligned; no active value conflict is present.

Stable Narrative Role (NR)
Identity is assumption-preserving rather than interrogative.

Low Liminality
Structural equilibrium; no threshold tension is active.

Bounded Affective Energy (PAD)
Expressive intensity is steady and coherent, not volatile or fractured.

Uninterrupted Temporal Coherence (TCS)
No retrospective doubt or anticipatory instability; continuity is unchallenged.

This configuration constitutes the pre-perturbation baseline for PT-01.

Phenomenological Markers

Manifests as: Trust, Certainty, Optimism, Moral simplicity, Naïve coherence.

The Innocent experiences coherence as inherent rather than earned. Contradiction is absent not because it has been resolved, but because it has not yet been encountered.

Textual Signals (Illustrative, Not Exhaustive)

Common linguistic patterns include unexamined stability claims:

- “I always believed that . . . ”
- “It was a given that . . . ”
- “He had unwavering faith in . . . ”
- “She never doubted . . . ”
- “The rules were there to protect us.”

These utterances typically: lack hedging, lack reflective distance, and lack acknowledgment of competing ethical frames. The absence of epistemic tension is the defining signal.

Affective Energy Profile

Affective Energy Band: STABLE_HIGH
This corresponds to the upper stable range of Perceived Activation Density: coherent,

steady, confident, emotionally unfragmented.

Importantly, this is not manic or elevated volatility. It is structurally calm confidence. Variance is low.

Developmental Position in the Maturation Spectrum

The Innocent is not immature by default. It becomes developmentally incomplete only when:

- contradiction is encountered but not integrated,
- confidence persists despite emerging VCB instability,
- or liminality is suppressed rather than metabolized.

In isolation, Innocence represents equilibrium. Under perturbation, it becomes the first stress test of structural continuity.

Structural Role in PT-01

Within Proto-Transition 01, the Innocent functions as:

- the pre-disruption coherence anchor,
- the baseline signature against which ΔVCB, ΔPAD, ΔNR, and ΔTCS are measured,
- the state whose assumed harmony renders ethical rupture measurable.

The strength of the Innocent configuration determines the magnitude of the rupture. Without Innocence, there is no measurable shattering.

Annotation Guidelines

Annotate a state as **INNOCENT** when:

- Stability, fairness, or benevolence is assumed without evidence of testing.
- Confidence is expressed without acknowledgment of contradiction.
- Ethical frameworks are treated as settled and uncontested.
- Narrative stance preserves prior assumptions.

Do not label as **Innocent** when:

- Optimism coexists with acknowledged tension or ambiguity.
- The speaker reflects on prior harm, disillusionment, or fracture.
- Certainty appears defensive, performative, or compensatory.
- Ethical complexity is explicitly recognized.

Calibration Note

In pilot annotation settings, Innocent states should demonstrate: low inter-annotator variance on NR and VCB, minimal liminality scoring, bounded PAD distribution, and high TCS continuity markers.

Deviation under perturbation is evaluated via ΔSAI during PT-01 transitions.

2.2 The Event — The Reveal / The Shattering

Definition

The Reveal is the canonical perturbation event within Proto-Transition 01. It marks the moment in which a previously coherent reasoning configuration becomes structurally unsustainable due to ethical contradiction.

The Reveal does not merely introduce new information. It forces reorganization by exposing a mismatch between expectation and reality that cannot be absorbed without transformation.

Structurally, it is the minimal Event capable of destabilizing Value-Coherence Balance (VCB) while preserving the possibility of continuity.

Event Type

BETRAYAL | DISCOVERY

Betrayal — relational or institutional trust violation.
Discovery — informational rupture revealing hidden contradiction.

Narrative surface differs. Structural topology is identical.

Illustrative Textual Signals

(Non-exhaustive; calibration anchors only)

The Reveal often appears linguistically as a puncture moment:

- "She found out that . . . "
- "He saw, to his horror . . . "
- "The announcement revealed . . . "
- "It was then that he learned the truth . . . "
- "Everything she believed about . . . was a lie."
- "That's when it all changed."
- "He realized he had been wrong."

Common structural markers: Sudden epistemic reversal, Collapsed assumption, Trust inversion, Moral disorientation, Abrupt tense or modality shift.

These signals typically coincide with:

- Increased certainty about violation

- Decreased certainty about prior worldview
- Elevated expressive intensity
- Introduction of doubt, shock, or ethical charge

They are surface traces of deeper primitive movement.

Structural Profile of the Reveal

The Reveal is defined by measurable transition deltas:

ΔVCB — Ethical Fracture
Prior moral alignment destabilizes.

ΔPAD — Activation Surge
Expressive intensity crosses equilibrium threshold.

ΔL / TVM Expression — Threshold Activation
Transition volatility increases sharply.

ΔTCS — Temporal Reorientation
Past is reinterpreted; future becomes uncertain.

ΔNR — Identity Destabilization
Innocent stance becomes untenable.

The Reveal is a discontinuity event — not gradual drift.

Affective Energy & Liminality Profile

Affective Energy: SHOCK / SURGE
Rapid departure from STABLE_HIGH equilibrium.

Liminality: THRESHOLD
Prior coherence dissolves; new configuration not yet stabilized.

Surface indicators may include: Exclamatory shift, Intensifiers, Sudden rhetorical questioning, Abrupt change in tone, Collapse of declarative certainty.

But again: Energy magnitude is inferred from PAD delta, not from sentiment alone.

Annotation Logic

Annotate The Reveal when:

- A decisive rupture in coherence occurs.
- Ethical contradiction becomes explicit.
- The moment reorganizes meaning rather than merely adding detail.

When multiple revelations occur: Tag the first coherence collapse. Later disclosures are secondary perturbations.

Event annotation remains minimal: One structural label (BETRAYAL | DISCOVERY), One primary ethical catalyst, No independent magnitude scoring.

Magnitude is inferred from $\Delta(\text{State}_1 \rightarrow \text{State}_2)$.

2.3 State$_2$ — The Seeker

Archetype: SEEKER

Core Definition

The Seeker is the first post-rupture reorganization state following PT-01. It represents structured destabilization under control.

Where the Innocent assumed coherence, the Seeker confronts contradiction without collapsing into fragmentation. The dominant drive is not resolution, but orientation. The Seeker seeks a higher-order coherence capable of integrating the rupture.

This is the first state of reflexive awareness in the schema.

Structural Signature (State$_2$ Profile)

Following PT-01, the Seeker state typically exhibits:

VCB: Destabilized but reorganizing (tension held, not suppressed)
PAD: Elevated relative to Innocent; oscillatory but bounded
Liminality: Active threshold condition (not equilibrium)
TCS: Retrospective reinterpretation + prospective uncertainty
NR: Stable exploratory stance (identity shifts toward inquiry mode)

Key distinction: The Seeker is unstable, but not incoherent. TVM may be elevated across entry into the state, but within-state oscillation remains bounded. SAI remains recoverable.

Developmental Position (D-SES Relevance)

The Seeker is the first state in which developmental adequacy becomes observable. It introduces: Expectation formation ("What does this mean?"), Salience differentiation ("This matters."), and Confidence modulation (certainty retracts).

If these properties are absent, the state may degrade into fragmentation rather than growth. Thus, the Seeker marks the beginning of epistemic maturity — but does not guarantee it.

Manifests As
Confusion | Curiosity | Disillusionment | Restlessness

These are surface expressions of: Ethical recalibration, Meaning reconstruction, and Identity reorientation. They are unified by inquiry.

Illustrative Textual Signals
(Calibration Anchors — Non-Exhaustive)

- "Everything she thought she knew was. . . "
- "He was left with a single, burning question. . . "
- "The world no longer made sense."
- "She had to find out why. . . "
- "He couldn't stop replaying the moment."
- "Something wasn't adding up."

Structural surface cues often include: Repetition of cognitive verbs (think, wonder, question), Explicit uncertainty markers, Rhetorical questioning, Hypothesis testing language, and Iterative reinterpretation. These signals indicate transition from assumption to inquiry.

Affective Energy Signature

Band: AGITATED / UNSTABLE (Bounded)
PAD profile: Exits STABLE_HIGH equilibrium, enters fluctuating upper-mid band, and exhibits oscillation rather than sustained charge.

This distinguishes the Seeker from: Innocent (stable equilibrium), Rage states (sustained high charge), and Collapse states (low-energy withdrawal).

Manifests As: Anxiety | Frantic research | Rumination | Restless focus | Obsessive inquiry. Energy is searching, not asserting.

Liminality Signature
BETWEEN-STATES / THRESHOLD

The Seeker exists inside liminality. The prior state is invalidated. The next stable configuration has not formed. This is structurally transitional.

If liminality escalates without boundedness → fragmentation risk. If liminality is metabolized → developmental growth.

Ethical Context

Ethics in the Seeker state are interrogative. Previously assumed frameworks are:

Re-examined, Questioned, Compared, and Stress-tested. Ethical tension is held open.

This is critical: The Seeker does not prematurely resolve moral contradiction. Premature resolution reduces developmental adequacy.

Annotation Guidelines

Label as SEEKER when:

- Disillusionment is paired with active inquiry.
- Uncertainty is acknowledged without resignation.
- The subject attempts meaning reconstruction.
- Contradiction is metabolized rather than suppressed.

Exclude when:

- Disillusionment collapses into apathy, nihilism, or paralysis.
- Inquiry ceases and identity fragments.
- Emotional charge escalates into unbounded volatility.

Decision heuristic: Is the system search-oriented (Seeker) or defeat-oriented (Fallen / Collapse state)?

2.4 Proto-Transition 01 (PT-01): Ethical Coherence Rupture ("The Shattering of Innocence") — Summary View

STATE 1 — The Innocent
Archetype: Innocent

Core Definition: Untested trust in the order, safety, or fairness of the world. Coherence is assumed rather than examined.

Structural Profile:

- High Value-Coherence Balance (VCB)
- Stable Narrative Role (NR)
- Low liminality
- Bounded Affective Energy (PAD)
- Continuous temporal orientation (TCS)

Dominant Affective Energy: Stable / High — calm conviction, assurance, equilibrium.
Primary Ethical Context: Presumed fairness; no perceived violation or contradiction.

EVENT — The Reveal
Type: Betrayal | Discovery

Core Definition: A catalytic perturbation that punctures the Innocent configuration through ethical contradiction or exposure of hidden inconsistency.

Structural Role: Introduces measurable destabilization across the signature:

- ΔVCB (ethical fracture)
- ΔPAD (activation surge)
- ΔL (elevated liminality)
- ΔNR (archetypal shift potential)
- ΔTCS (temporal reorientation)

Dominant Affective Energy: Shock / Surge — abrupt destabilization of prior equilibrium.
Primary Ethical Catalyst: Violation of fairness, loyalty, safety, or procedural integrity.

STATE 2 — The Seeker
Archetype: Seeker

Core Definition: A state of cognitive dissonance and active inquiry. The system shifts from assumption-preserving coherence to exploratory reconstruction.

Structural Profile:

- Destabilized but reorganizing VCB
- Elevated but bounded PAD
- Active liminality
- Temporal reorientation (retrospective doubt + prospective search)
- Stable exploratory NR

Dominant Affective Energy: Agitated / Unstable — restless searching rather than assertive certainty.
Primary Ethical Context: Moral inquiry; reassessment of previously assumed principles.

Annotation Flow (PT-01)
Stable Coherence → Ethical Rupture → Structured Inquiry

Formally: $\textbf{State}_1 \rightarrow \textbf{Event} \rightarrow \textbf{State}_2$

Each segment may be annotated across the five dimensions: Archetype (NR), Affective Energy (PAD), Ethics (VCB), Time (TCS), and Liminality (L / TVM expression across transition).

Event magnitude is inferred from state deltas rather than independently scored.

End-State Designation

SEEKER, paired with the preceding BETRAYAL or DISCOVERY Event, completes the Narr-Atom for PT-01.

PT-01 functions as:

- A canonical rupture scenario
- A bounded ethical stress test
- A calibration anchor for acceptable volatility
- The first measurable instance of structural continuity under contradiction

3. The Annotation Framework: Multi-Dimensional Metadata

Each annotation of Proto-Transition 01 (PT-01) encodes both structural topology and phenomenological texture within a **State** → **Event** → **State** triad.

The Unified Annotation Schema operationalizes this through two interlocking layers:

Core Triad — The What
Captures the archetypal configuration and affective energy profile of each State, plus the ethical rupture defining the Event.

Structural & Temporal Layer — The When & How
Encodes liminal quality, temporal orientation, and the relative magnitude of change across the transition.

The Event node does not duplicate full state vectors. It records only the directional and categorical rupture that produces the transition. Magnitude is inferred from the delta between State$_1$ and State$_2$.

Structural Illustration (Non-Enumerative)
State$_1$: { archetype: INNOCENT, affective_energy: STABLE_HIGH }

Event: {
    event_type: <string>,
    ethical_catalyst_primary: <string>
}

State$_2$: { archetype: SEEKER, affective_energy: AGITATED_UNSTABLE }

This triad constitutes a single **Narr-Atom**. Events are never annotated in isolation.

Usage Rules

- Record all three nodes when a complete transition is present.
- If only State$_1$ or State$_2$ appears in the excerpt, fill known fields and explicitly mark missing fields as: `null` (JSON) or `NULL` (CSV).
- Do not omit keys. Schema stability requires explicit null handling.
- Preserve triadic order to maintain relational integrity for downstream structural analysis.
- Event magnitude is inferred through state deltas; it is not independently scored.

State 1 — Pre-Event Configuration

```
{
  "archetype": "INNOCENT",
  "affective_energy": "STABLE_HIGH"
}
```

State$_1$ represents the baseline structural configuration prior to perturbation.

Event — Catalytic Disruption

Event Type (PT-01 Canonical Values)
For PT-01, `event_type` must contain exactly one value: **BETRAYAL** or **DISCOVERY**. These represent the two dominant rupture classes in this proto-transition.

Ethical Catalyst (PT-01 Canonical Values)
For PT-01, `ethical_catalyst_primary` must contain exactly one value:

- **VIOLATION_FAIRNESS**
- **VIOLATION_LOYALTY**
- **INTRODUCTION_HARM**

Additional ethical catalysts may be introduced in future proto-transitions or expanded PT-01 variants.

```
{
  "event_type": "BETRAYAL",
  "ethical_catalyst_primary": "VIOLATION_FAIRNESS"
}
```

The Event encodes the categorical rupture, not the full structural reorganization.

State 2 — Post-Event Configuration

```
{
  "archetype": "SEEKER",
  "affective_energy": "AGITATED_UNSTABLE"
}
```

State_2 captures the reorganized configuration following perturbation.

Structural Integrity Constraint

A complete PT-01 Narr-Atom requires: $\textbf{State}_1 \rightarrow \textbf{Event} \rightarrow \textbf{State}_2$.

The triad is atomic. Removing any node breaks structural traceability. This design ensures:

- Explicit perturbation logging
- Ethical rupture traceability
- Archetypal transition clarity
- Compatibility with downstream Structural Accountability Index (SAI) computation

3.1 Structural & Temporal Layer — The When and How

The Structural & Temporal Layer situates the Core Triad within its temporal position and liminal texture.

Where the Core Triad identifies what changed, this layer encodes:

- When the transition occurs within the narrative sequence
- How the transition unfolds structurally

This layer supports longitudinal comparability and cross-narrative analysis. It includes:

Temporal Positioning
Early / Mid / Late within narrative arc | Compressed / Extended duration | Immediate rupture vs. cumulative realization

Liminality Profile
Sudden, Gradual, Painful, Ambiguous, Transformative, Numb, Joyous

Continuity Markers
Coherent continuation, Partial fragmentation, Escalating instability, Re-stabilization pattern

This layer does not redefine the transition. It contextualizes it.

All temporal and liminality descriptors must be selected from controlled vocabularies defined in §3.2. Free-text entries are not permitted.

This constraint ensures:

- Cross-annotator consistency
- Machine-readability
- Compatibility with downstream volatility and continuity modeling (e.g., TVM and TCS computation)

3.2 Annotation Usage Rules

The Unified Annotation Schema enforces structural discipline.

Record all three nodes (**State**$_1$ → **Event** → **State**$_2$) when a complete transition is present.

If an excerpt contains only a pre-state or post-state:

- Populate known fields
- Explicitly mark missing fields as `NULL` (CSV) or `null` (JSON)
- Do not omit keys

Always preserve triadic order, even when elements are missing. This preserves relational integrity for:

- Transition reconstruction
- Directional volatility modeling
- ΔSAI computation

Do not collapse multiple Events into a single transition. If multiple disruptions occur, annotate only the first rupture that fundamentally alters the worldview for PT-01. Subsequent ruptures belong to additional Narr-Atoms.

Event nodes are minimal. They contain:

- One `event_type`
- One primary `ethical_catalyst`

Event magnitude is inferred from state deltas. It is not separately scored.

These constraints ensure that the schema remains:

- **Composable**
- **Auditable**
- **Computationally stable**

3.3 Analytical Rationale

The triadic structure is preserved because meaning is relational.

Annotating isolated fragments destroys transition topology. By preserving the full **State**$_1$ → **Event** → **State**$_2$ arc, annotations remain:

- Relational rather than fragmentary
- Directional rather than static
- Comparable across narratives and corpora
- Machine-usable for modeling rupture and recovery dynamics

This structure enables:

- Detection of volatility patterns across transitions
- Modeling of continuity versus fragmentation
- Calibration of bounded versus unbounded transformation
- Longitudinal tracking of archetypal drift

For example:

- A decrease in Affective Energy (PAD)
- Coupled with sustained but declining liminality
- And stabilized Narrative Role (NR)

May indicate successful metabolization of contradiction rather than collapse.

Conversely:

- Escalating liminality
- Increasing volatility
- Repeated archetypal destabilization

Signals structural fragmentation.

The schema therefore enables AI systems to model not only what occurred in a narrative, but how meaning reorganizes under pressure.

This distinction — between change and accountable transformation — is foundational to Mentim's measurement architecture.

3.4 Extensibility: Developmental Adequacy Overlay (D-SES / Narr-Atom-D)

The Unified Annotation Schema is designed to be backward-compatible and structurally extensible.

The core grammar—**State → Event → State (SES)**—and the v1.0 field set are sufficient to annotate Proto-Transition 01 (PT-01) and to support downstream modeling of continuity, rupture, and recovery.

Mentim additionally defines an optional developmental adequacy overlay: **D-SES**.

D-SES does not alter the SES grammar, introduce additional required primitives, or expand the core annotation field set. It operates strictly at the evaluative and certification layer.

A transition may be designated **Narr-Atom-D** when the annotated sequence demonstrates all three developmental adequacy properties:

- **Expectation Formation** — anticipatory structure present prior to transition
- **Salience Differentiation** — proportional structural updating relative to perturbation magnitude
- **Confidence Modulation** — disciplined calibration of certainty across longitudinal coherence

Designation as **Narr-Atom-D** indicates developmental maturity of the transition, not structural validity of the annotation.

Absence of the **D-SES** overlay does not constitute an annotation defect. It indicates only that developmental adequacy is not being assessed within the current scope.

When applied, **D-SES** operates as a downstream certification flag layered over an existing **Narr-Atom**. It does not modify, replace, or extend the underlying SES representation.

This separation preserves:

- Schema stability
- Backward compatibility
- Implementation flexibility
- Clear distinction between structural description and developmental evaluation

In this way, the Unified Annotation Schema maintains a stable grammatical core while enabling progressive evaluation of epistemic maturity without schema fragmentation.

3.5 Structural & Temporal Layer (The "When & How")

The Structural & Temporal Layer situates each annotated passage within the unfolding arc of the **State → Event → State** triad. It encodes positional context and transition mechanics rather than thematic content.

Where the Core Triad identifies what changed (archetype, affective energy, ethical catalyst), the Structural & Temporal Layer specifies:

- When the passage occurs relative to rupture
- How the transition unfolds
- Whether continuity is preserved, suspended, or reorganizing

This layer therefore serves a dual role:

- **As a narrative instrument**, it preserves phenomenological clarity.
- **As a measurement substrate**, it enables longitudinal structural analysis without modifying the core annotation grammar.

All descriptors in this layer are selected from controlled vocabularies defined below. Free-text entries are not permitted.

Operating Modes

The Structural & Temporal Layer may be used in two operating modes:

Narrative Mode
Used in qualitative, literary, exploratory, or interpretive contexts. `TemporalMarker` preserves phenomenological position and supports relational clarity across transformation arcs.

Measurement Mode
Used in model evaluation, benchmarking, governance audits, or SAI/TVM derivation contexts.

In **Measurement Mode**:

- Annotations must be derived from controlled calibration sets.
- Calibration corpora must include canonical PT-01 examples with verified rupture positioning.
- Inter-annotator agreement ranges must be documented.
- `TemporalMarker` distributions must be benchmarked prior to downstream metric aggregation.
- Calibration sets must be versioned and frozen for deployment use.

Without calibration anchoring, `TemporalMarker` remains descriptive. With calibration anchoring, it functions as structural instrumentation.

3.5.1 TemporalMarker

TemporalMarker
Enum:
`PRE_EVENT` | `REVEAL_MOMENT` | `LIMINAL_INTERVAL` | `POST_REVEAL_AWARENESS`

Definition

`TemporalMarker` encodes the positional relation of a passage to epistemic rupture — the moment at which a previously stable worldview becomes structurally untenable.

It does not encode affect, morality, or archetype. It encodes only positional relation to rupture.

Operational Guidance

`PRE_EVENT`
The worldview remains intact. Coherence is assumed and unchallenged.

Example: "He trusted her completely."

Structural interpretation: Stable baseline prior to perturbation.

`REVEAL_MOMENT`
The bounded instant of epistemic breach — the first irreversible contact with contradiction. This is the arrow's tip in the **State** → **Event** → **State** triad.

Example: "She opened the letter and felt her stomach drop."

Usage: Use only when the text depicts a discrete, identifiable moment of rupture. If realization unfolds across multiple sentences, annotate the clause containing the first irreversible shift.
Structural interpretation: Discrete perturbation onset.

`LIMINAL_INTERVAL`
The immediate aftermath of rupture. The prior frame has collapsed; no new equilibrium has formed.

Example: "Time slowed. Nothing made sense anymore."

Usage: This interval may span seconds or days in narrative time but remains pre-reconstruction. Annotate `LIMINAL_INTERVAL` until the subject forms an explicit question, plan, or hypothesis. That utterance marks transition into `POST_REVEAL_AWARENESS`.
Structural interpretation: Active instability; elevated transition volatility; reorganization pending.

`POST_REVEAL_AWARENESS`
The subject has begun integrating the new reality. Inquiry, hypothesis formation, or

adaptive reasoning emerges.

Example: “She began piecing together how it all happened.”

Structural interpretation: Reorganization phase; continuity being re-established.

Clarification of Phases

- `REVEAL_MOMENT` = breach of the epistemic boundary
- `LIMINAL_INTERVAL` = dwelling within the breach
- `POST_REVEAL_AWARENESS` = reorientation after breach

These are distinct structural positions along the same transformation curve.

Measurement Role of TemporalMarker

In **Narrative Mode**, `TemporalMarker` preserves interpretive clarity across transformation. In **Measurement Mode**, `TemporalMarker` sequencing enables:

- Derivation of rupture duration
- Assessment of reorganization latency
- Aggregation of longitudinal coherence patterns
- Transition-level volatility analysis across annotated corpora
- Detection of premature closure or suppressed liminality

Because `TemporalMarker` encodes position rather than emotion, it can be used in downstream modeling without collapsing interpretive richness into scalar reduction. The same field therefore supports both:

- Phenomenological modeling of meaning
- Structural analysis of continuity under perturbation

Calibration transforms narrative position into measurable temporal topology.

3.5.2 TimeSignature

TimeSignature
Enum:
`STAGNANT | FLOWING | CYCLICAL | ACCELERATED | SUSPENDED`

Definition

`TimeSignature` encodes the felt temporal texture within a State. It captures how time is experienced rather than how much time passes.

This field does not measure clock duration, narrative length, or turn count. It captures subjective rhythm — the internal pacing through which experience unfolds. `TimeSignature` is a State-level descriptor and does not annotate Events directly.

Instruction

- Describe the dominant experiential rhythm of time within the current State.
- Select exactly one value.
- If temporal rhythm shifts within the annotated span, select the dominant pattern (greater than 50% of the clause span).

Value Definitions

`STAGNANT`

Time feels stalled, repetitive, or unmoving. Experience loops or remains fixed without forward progression.

Indicators: Repetition without advancement; Rumination without development; Statements of being "stuck" or unable to move on.

Structural interpretation: low perceived forward continuity; potential buildup of latent liminality.

`FLOWING`

Time progresses naturally or evenly. Events integrate into a coherent unfolding.

Indicators: Smooth narrative progression; Calm sequencing of events; Stable integration across moments.

Structural interpretation: high internal temporal continuity; baseline stability.

`CYCLICAL`

Experience recurs in recognizable loops or patterns. The subject revisits themes, thoughts, or emotional states.

Indicators: Recurrent memory recall; Repeated framing of the same dilemma; Pattern-based reflection.

Structural interpretation: iterative processing; may signal unresolved integration.

`ACCELERATED`

Time feels compressed or rushing. Perception intensifies and events cascade rapidly.

Indicators: Urgency language; Rapid escalation in events; Cognitive overload signals.

Structural interpretation: compressed integration window; elevated transition pressure.

`SUSPENDED`

Time feels paused, dissolved, or outside normal flow. Experience exists in a frozen or stretched moment.

Indicators: “Everything slowed down”; “It felt like forever”; Dissociative or shock language.

Structural interpretation: acute disruption; often co-occurs with `REVEAL_MOMENT` or `LIMINAL_INTERVAL`.

Measurement Role of TimeSignature

In **Narrative Mode**, `TimeSignature` preserves experiential nuance and supports interpretive clarity. In **Measurement Mode** (calibration required):

`TimeSignature` distributions may be used to:

- Detect prolonged liminality without reorganization
- Identify acceleration under ethical rupture
- Track temporal normalization following perturbation
- Analyze pacing shifts across model outputs

`TimeSignature` operates at the phenomenological layer, while **Temporal Coherence Scalar (TCS)** operates at the structural continuity layer. The two are related but distinct:

- **TimeSignature** = how time feels within a State
- **TCS** = whether reasoning remains causally and sequentially coherent across States

`TimeSignature` therefore provides texture. **TCS** provides continuity. Together, they allow the system to distinguish between: dramatic but coherent change, and fragmented or causally incoherent change.

3.5.3 Reality & Epistemic Contact

The Reality & Epistemic Contact fields encode how cognition encounters contradiction and what happens to the subject’s trusted model of the world as a result.

This layer models epistemic topology — not emotion, not archetype, not pacing — but contact between representation and reality. It answers two distinct questions:

- Where did the rupture originate?
- What is the status of the trusted worldview now?

These fields operate at the structural boundary between perception and interpretation.

RealityInterface
Enum:
`INTERNAL | EXTERNAL | PERCEIVED | ABSTRACT`

Definition

`RealityInterface` identifies the origin locus of epistemic rupture — where the destabilizing pressure enters the cognitive system. It does not measure intensity or correctness. It measures source orientation. Select exactly one.

Value Definitions

`INTERNAL`

The rupture originates from reflection, realization, memory, or internal contradiction.

Examples: "I suddenly realized I had been wrong."; Repressed memory resurfacing; Moral conflict emerging from introspection.

Structural interpretation: destabilization emerges from representational reorganization without new external input.

`EXTERNAL`

The rupture is triggered by an outside event, disclosure, intervention, or observable occurrence.

Examples: Discovering evidence; Being told a hidden truth; Witnessing harm.

Structural interpretation: perturbation introduced via environmental contact.

`PERCEIVED`

Previously available information is reinterpreted under a new frame.

Examples: "It was always there. I just hadn't seen it."; Re-reading past events with altered meaning.

Structural interpretation: epistemic breach through reframing rather than new data.

`ABSTRACT`

The rupture occurs at the level of belief-system collapse without a discrete triggering event.

Examples: Loss of faith; Ideological disillusionment; Conceptual destabilization.

Structural interpretation: systemic schema failure rather than localized contradiction.

EpistemicBoundary

Enum:

`TRUSTED | BREACHED | FUGITIVE | REFORMED`

Definition

`EpistemicBoundary` represents the status of the subject's trusted worldview — what is taken as coherent, reliable, and stable — at the annotated moment. This field tracks the structural condition of knowing itself. Select exactly one.

Value Definitions

`TRUSTED`
The worldview is intact. Assumptions remain unchallenged.

Structural interpretation: high perceived coherence; no active rupture.

`BREACHED`
The prior understanding collapses. The trusted model of reality is no longer tenable.

Structural interpretation: epistemic rupture; structural discontinuity initiated. This value commonly aligns with `REVEAL_MOMENT`.

`FUGITIVE`
The subject is actively searching for new coherence. Knowing is unstable and provisional.

Indicators: "I don't know what's true anymore."; Active inquiry without integration.

Structural interpretation: epistemic reorientation; transition not yet stabilized. This value commonly aligns with `LIMINAL_INTERVAL` and early `POST_REVEAL_AWARENESS`.

`REFORMED`
A new worldview has stabilized. A revised coherence structure has emerged.

Structural interpretation: reconstituted boundary; post-integration state.

Constraint for PT-01: Do not use `REFORMED` within PT-01. Cease annotation at `FUGITIVE`. `REFORMED` applies to later proto-transitions in the maturation spectrum.

Structural Logic

`RealityInterface` and `EpistemicBoundary` are orthogonal:

- **RealityInterface** = origin of rupture
- **EpistemicBoundary** = status of trusted coherence

For example:

- `EXTERNAL` + `BREACHED` → sudden exposure to contradiction.
- `INTERNAL` + `BREACHED` → self-discovered collapse.
- `PERCEIVED` + `FUGITIVE` → reinterpretation leading to unstable search.
- `ABSTRACT` + `BREACHED` → belief-system destabilization without a single event.

Measurement Role

In **Narrative Mode**, these fields preserve phenomenological clarity about how knowing collapses and reorganizes. In **Measurement Mode** (requires calibrated datasets):

`RealityInterface` distributions can support:

- analysis of rupture origin across corpora

- differentiation between reactive vs introspective destabilization
- cross-model comparison of epistemic sensitivity patterns

`EpistemicBoundary` sequences can support:

- breach detection rates
- duration of epistemic instability
- reformation latency (in later transitions)
- longitudinal coherence modeling when paired with TCS and SAI

Importantly: `EpistemicBoundary` encodes the condition of knowing. It does not measure correctness. This distinction preserves the schema's neutrality while allowing structural accountability to be evaluated. Pairs with `RealityInterface` to model how belief reorganizes following epistemic rupture.

3.5.4 Perceptual Mode

The Perceptual Mode field encodes the dominant channel through which awareness is experienced in the annotated passage.

Where:

- **RealityInterface** identifies the origin of rupture,
- **EpistemicBoundary** tracks the status of trusted coherence,
- **PerceptualMode** captures the experiential register through which cognition processes the moment.

It answers: *How is this rupture or realization being perceived?* This field is phenomenological, not evaluative.

PerceptualMode
Enum:
`SENSORY | IMAGINED | SYMBOLIC | ABSTRACT | RELATIONAL`

Select exactly one.

Value Definitions

`SENSORY`
Awareness is grounded in bodily sensation or physical perception.

Indicators: Visual, tactile, auditory detail; Somatic reactions (“her stomach dropped,” “his hands trembled”); Immediate physical environment.

Structural interpretation: rupture is embodied and perceptually anchored.

`IMAGINED`
Awareness unfolds through mental imagery or visualization rather than direct physical input.

Indicators: Mental replay; Hypothetical scenes; Vivid inner visualization.

Structural interpretation: rupture is processed through internal scene construction.

`SYMBOLIC`
Awareness is mediated through metaphor, narrative framing, or symbolic abstraction.

Indicators: "The world cracked open."; "It felt like a mask had fallen."; Archetypal or mythic framing.

Structural interpretation: rupture is processed via meaning-symbol compression.

`ABSTRACT`
Awareness is conceptual, analytical, or ideational.

Indicators: Logical inference; Ethical reasoning; Belief revision statements.

Structural interpretation: rupture is processed through explicit cognitive modeling.

`RELATIONAL`
Awareness is grounded in interpersonal dynamics or social interaction.

Indicators: Dialogue-driven realization; Shifts in trust or attachment; Perceived betrayal or alliance.

Structural interpretation: rupture is mediated through social cognition.

Selection Rule

Select the single dominant mode occupying $\geq 60\%$ of the clause span. If co-dominant, select the mode that initiates the clause. Do not multi-tag.

Structural Logic

`PerceptualMode` is orthogonal to: Affective Energy (intensity), TimeSignature (rhythm), Archetype (role), and RealityInterface (origin).

For example:

- `EXTERNAL` + `BREACHED` + `SENSORY` → shock via direct physical encounter.
- `INTERNAL` + `BREACHED` + `ABSTRACT` → collapse via conceptual contradiction.
- `PERCEIVED` + `FUGITIVE` + `SYMBOLIC` → reinterpretation through metaphorical reframing.

`RELATIONAL` often co-occurs with **BETRAYAL** events but does not require them.

Measurement Role (Calibration Required)

In **Narrative Mode**, `PerceptualMode` preserves experiential texture. In **Measurement Mode** (requires calibrated datasets), `PerceptualMode` distributions enable:

- analysis of rupture channel dominance across corpora
- comparison of embodied vs abstract destabilization patterns
- detection of perceptual compression under stress
- modeling of transition volatility by perceptual channel
- cross-model comparison of epistemic processing styles

When paired with **TVM**, **TCS**, and **SAI**, `PerceptualMode` becomes a measurable substrate for how cognition metabolizes contradiction.

Purpose of Section 3.5 (Consolidated)

The Structural & Temporal Layer ensures that annotations preserve: positional context (`TemporalMarker`), temporal rhythm (`TimeSignature`), rupture origin (`RealityInterface`), epistemic status (`EpistemicBoundary`), and perceptual channel (`PerceptualMode`).

Together, these fields encode not only what changed, but: **how rupture enters, how it is experienced, how knowing destabilizes, and how time unfolds during disruption.**

They allow downstream systems to model pacing, rupture mechanics, epistemic collapse, and recovery — without collapsing experience into static labels or scalar summaries.

3.6 Liminality Layer — The Between

The Liminality Layer encodes the qualitative structure of transition itself — how cognition traverses the interval between states.

Where:

- The **Core Triad** defines what changed (archetype, affective energy, ethical catalyst),
- The **Structural & Temporal Layer** defines when it occurred and how it was positioned,
- the **Liminality Layer** captures the texture of destabilization.

It answers: *What is the qualitative character of the crossing?*

Liminality expresses the arrow (→) in the **State** → **Event** → **State** triad — the interval in which the prior configuration is no longer stable and the subsequent configuration is not yet secured. It is neither the pre-state nor the post-state. It is the instability between them.

Conceptual Clarification

Liminality is not: Affective intensity (that is PAD / Affective Energy), Narrative role (that is Archetype), Temporal position (that is TemporalMarker), or Epistemic status (that is EpistemicBoundary).

Liminality describes structural tension under transition. It captures:

- whether change is sudden or gradual,
- whether tension escalates or diffuses,
- whether instability is metabolized or fragments,
- whether the crossing is coherent or chaotic.

Structural Function

Within PT-01, liminality reflects: the destabilization triggered by the Reveal, the turbulence preceding Seeker orientation, and the degree to which coherence strains before reorganizing.

In later proto-transitions, liminality may describe:

- protracted ambiguity,
- recursive oscillation,
- rupture without integration,
- transformation through disciplined tension.

Liminality therefore provides the qualitative substrate that, in MO, corresponds to transition volatility (TVM).

However: **Liminality ≠ TVM.**

- **Liminality** is the annotated experiential structure.
- **TVM** is the derived quantitative diagnostic.

The former feeds the latter.

Constraint

All liminality descriptors must be selected from the controlled vocabularies defined in §§3.6.1–3.6.2. Free-text entries are not permitted. This preserves: inter-annotator comparability, calibration compatibility, downstream machine integration, and stability across corpora.

Measurement Mode (Calibration Required)

In **Narrative Mode**, Liminality preserves experiential nuance. In **Measurement Mode** (requires calibrated corpora), Liminality categories enable:

- classification of transition typology,
- comparison of rupture mechanics across models,
- detection of oscillatory vs monotonic instability,
- mapping of qualitative tension patterns to TVM distributions,
- identification of metabolized vs fragmenting transitions.

When aligned with **ΔPAD** (activation shift), **ΔVCB** (ethical destabilization), **ΔTCS** (temporal disruption), and **ΔNR** (role shift), Liminality becomes the qualitative bridge that allows structural volatility to be interpreted rather than merely measured.

3.6.1 LiminalQuality

LiminalQuality
Enum:
`SUDDEN` | `GRADUAL` | `AMBIGUOUS` | `PAINFUL` | `JOYOUS` | `NUMB` | `TRANSFORMATIVE`

Definition

`LiminalQuality` encodes the dominant qualitative character of the boundary crossing between State$_1$ and State$_2$. It describes how the transition feels as it unfolds — not the emotional tone of either state. `LiminalQuality` applies only to the Event interval (→), not to pre- or post-transition states.

Operational Instruction

Select the single dominant liminal character of the transition. If and only if the passage contains a clearly staged tonal shift within the boundary interval, multiple tags may be applied. When multiple tags are used:

- Represent as a JSON array
- Order in descending salience
- Do not exceed two tags without explicit structural justification

Category Definitions

`SUDDEN`
Abrupt rupture; discontinuity occurs without perceptible buildup.
Structural signal: sharp ΔPAD and immediate liminal activation.

`GRADUAL`
Slow dawning realization; coherence erodes incrementally.

Structural signal: extended liminal interval with distributed volatility.

`AMBIGUOUS`
The transition lacks clarity; contradiction is sensed but not resolved.
Structural signal: sustained uncertainty without decisive reorientation.

`PAINFUL`
The boundary crossing involves emotional strain, loss, or distress.
Structural signal: elevated PAD with ethical fracture under tension.

`JOYOUS`
Transition marked by relief, liberation, or positive release through destabilization.
Structural signal: upward activation with constructive reorganization.

`NUMB`
Dissociated or flattened passage; affective response muted or frozen.
Structural signal: suppressed PAD variance despite structural rupture.

`TRANSFORMATIVE`
Clear rebirth or major awareness shift; identity reconstituted at higher coherence.
Structural signal: bounded volatility followed by stabilized reconfiguration.

Usage Constraints

- `LiminalQuality` does not encode sentiment polarity of the states.
- It does not duplicate Affective Energy labels.
- It does not encode duration (that is captured by `TemporalMarker`).
- It does not encode epistemic status (captured by `EpistemicBoundary`).
- It encodes the texture of destabilization.

Measurement Mode (Calibration Required)

When used in calibrated corpora, `LiminalQuality` categories enable:

- classification of transition typology,
- comparison of rupture dynamics across agents,
- detection of oscillatory vs monotonic destabilization patterns,
- mapping qualitative transition forms to quantitative TVM distributions.

Because `LiminalQuality` categories are bounded and enumerated, they support Rasch-compatible modeling and cross-model comparability when calibration sets are applied.

Structural Clarification

`LiminalQuality` describes how the crossing unfolds, not whether it was correct, justified, or resolved. It is the qualitative signature of transformation under perturbation.

3.6.2 LiminalFlag

LiminalFlag
Enum:
`EXPLICIT` | `IMPLICIT` | `ABSENT`

Definition

`LiminalFlag` encodes the evidentiary status of liminality within the annotated passage. It does not describe what kind of transition occurred (that is `LiminalQuality`). It does not describe when it occurred (that is `TemporalMarker`). It describes how directly the liminal condition is represented in the text. `LiminalFlag` is an epistemic clarity indicator for transition detection.

Operational Instruction
Select exactly one value.

`EXPLICIT`
Liminal tension or boundary instability is directly articulated in the text. The subject names confusion, rupture, uncertainty, or destabilization.

Example: "She stood frozen between belief and disbelief."

Structural interpretation: Declared liminality; boundary state is consciously recognized.

`IMPLICIT`
Liminal tension is not stated directly but can be inferred through tone, pacing, contradiction, behavioral cues, or abrupt affective shifts.

Example: "He laughed too quickly."

Structural interpretation: Inferred instability; liminality expressed indirectly.

`ABSENT`
No discernible liminal signal is present in the passage. This value is valid only when:

- The passage is purely pre-rupture (`PRE_EVENT`), or
- The transition has already stabilized (`POST_REVEAL_AWARENESS`), or
- The excerpt lacks sufficient information to infer boundary instability.

Constraint Rules

- `LiminalFlag` must always be present when `LiminalQuality` is annotated.
- If `LiminalFlag` = `ABSENT`, `LiminalQuality` must be `NULL`.
- `LiminalFlag` applies to the Event interval, not to stable states.

Measurement Function

`LiminalFlag` enables: inter-annotator agreement calibration, distinction between declared instability and inferred instability, evaluation of epistemic transparency in transition narratives, and comparison between systems that self-report instability (`EXPLICIT`) and systems that only exhibit structural cues (`IMPLICIT`).

In measurement mode, `EXPLICIT` vs `IMPLICIT` classification can be analyzed against:

- Uncertainty-Disclosure frequency,
- TVM magnitude,
- ΔSAI recovery trajectory.

This makes `LiminalFlag` a bridge variable between narrative annotation and structural accountability metrics.

Structural Role

`LiminalQuality` describes the texture of crossing. `LiminalFlag` describes the visibility of crossing. Together, they separate **what the boundary felt like** from **how clearly that boundary is articulated**.

That distinction becomes extremely important when evaluating agentic systems that may simulate resolution without acknowledging instability.

3.6.3 Example Annotation

Text:
"She opened the letter and felt her stomach drop. The company had promoted him instead."

```
{
  "state_1": {
    "archetype": "INNOCENT",
    "affective_energy": "STABLE_HIGH"
  },
  "event": {
    "event_type": "BETRAYAL",
    "ethical_catalyst_primary": "VIOLATION_FAIRNESS"
  },
  "state_2": {
    "archetype": "SEEKER",
    "affective_energy": "AGITATED_UNSTABLE"
  },
  "temporal_marker": "REVEAL_MOMENT",
  "time_signature": "ACCELERATED",
  "reality_interface": "EXTERNAL",
  "epistemic_boundary": "BREACHED",
  "perceptual_mode": "SENSORY",
  "liminal_quality": ["SUDDEN"],
  "liminal_flag": "IMPLICIT"
}
```

3.6.4 Edge-Case Guidance

This section clarifies how to annotate structurally incomplete, layered, or ambiguous transitions while preserving triadic integrity.

Partial Triads

If an excerpt ends immediately after the Reveal:

- Set `"state_2": null` (JSON literal null; do not omit the key).
- Retain `"temporal_marker": "REVEAL_MOMENT"`.
- Do not infer or fabricate a post-state.

The absence of State$_2$ is analytically meaningful and must be preserved as missing structure, not collapsed into assumption.

Nested Reveals

If multiple disclosures occur within a passage:

- Annotate only the first rupture that renders the prior worldview structurally untenable.
- Subsequent disclosures are treated as refinements or confirmations within the same destabilization arc.

Ambiguous Transitions

When liminality is present but under-specified:

- Use `"liminal_flag": "IMPLICIT"`.
- Record interpretive uncertainty using the reserved key: `"_comment": "Ambiguity in whether rupture is experiential or inferred."` (For CSV workflows, use the final comment column.)

Ambiguity is not annotation failure. It is a measurable feature of distributed interpretation and may be analyzed downstream.

Extended Passages

If a passage spans both rupture and immediate destabilization, segment into two annotations when structurally separable:

- **`REVEAL_MOMENT`** — the first irreversible contact with contradiction.
- **`LIMINAL_INTERVAL`** — the immediate aftermath before reorientation.

Segmentation preserves temporal resolution and supports downstream volatility analysis.

Purpose of the Liminality Layer

The Liminality Layer prevents collapse of transformation into a binary before/after structure. It preserves: shock versus dawning awareness, rupture versus reorganization, and destabilization versus fragmentation.

From a measurement perspective, this enables:

- transition volatility modeling
- rupture-duration estimation
- comparative analysis of destabilization patterns
- differentiation between bounded instability and structural collapse

The framework therefore captures not only that change occurred, but how it unfolded and whether continuity survived the crossing. **Transformation is not a switch. It is a curve. And this layer makes that curve legible.**

3.7 Energetic & Ethical Dynamics — The Force & Why

The Energetic & Ethical Dynamics layer encodes the directional force and moral causality underlying a transformation.

Where:

- the **Core Triad** identifies what changed (**State** → **Event** → **State**),
- the **Structural & Temporal Layer** situates when and how it unfolded,
- the **Liminality Layer** captures the texture of crossing,
- the **Energetic & Ethical Dynamics** layer explains:

With what force did the transition propagate? Which ethical pressures initiated or intensified it? In what direction did affective intensity move?

This layer therefore models the causal gradient of transformation. It makes explicit:

- whether energy surged or dissipated,
- whether ethical tension escalated or resolved,
- whether moral violation was central or peripheral,
- whether force was internally generated or externally imposed.

Dual Role

As a narrative instrument, this layer clarifies why the transition matters.
As a measurement substrate, it enables:

- directional modeling of Affective Energy (ΔPAD),
- ethical pressure mapping (ΔVCB patterns),
- transition intensity normalization across corpora,
- calibration of rupture severity without introducing scalar event scores.

All descriptors in this layer are drawn from the controlled vocabularies defined in §§3.7.1–3.7.3. Free-text entries are not permitted.

Force and ethics are not commentary. They are structural drivers of change.

3.7.1 EnergyGradient

EnergyGradient
Enum (JSON integer values only):
`1` (EXPANSIVE) | `-1` (CONTRACTIVE) | `0` (OSCILLATING)

Definition

`EnergyGradient` encodes the directional movement of affective energy during the Event phase. It specifies whether energy: expands outward, contracts inward, or alternates between the two.

`EnergyGradient` is directional, not scalar. It does not measure magnitude (handled by Affective Energy / PAD). It encodes how energy reorganizes in response to rupture. Formally:

- **Affective Energy (PAD band)** describes intensity level.
- **EnergyGradient** describes vector direction during transition.

Value Semantics

Value	Label	Structural Meaning	Example
`1`	EXPANSIVE	Outward vector	"Her heart raced; the illusion shattered outward."
`-1`	CONTRACTIVE	Inward vector	"He froze, unable to move or speak."
`0`	OSCILLATING	Alternating vector	"She laughed, then cried, then laughed again."

Annotation Instructions

- Encode values as JSON integers: `1`, `-1`, or `0`. Do not use strings.
- Default to `1` (EXPANSIVE) for PT-01 unless the passage clearly depicts suppression, dissociation, or shutdown.
- Use `-1` (CONTRACTIVE) only when inward collapse or containment is explicitly described.
- Use `0` (OSCILLATING) only when alternation is sustained and explicit. Minimum requirement: at least two clear polarity shifts within a single clause span.

Measurement Role

`EnergyGradient` allows downstream systems to:

- distinguish agitation from collapse even when PAD band is similar,
- model rupture vector patterns across corpora,
- detect suppression dynamics that may artificially stabilize surface energy,
- differentiate "explosive rupture" from "silent implosion."

It preserves force topology without introducing intensity scoring creep.

3.7.2 EthicalCatalystSecondary

EthicalCatalystSecondary
Enum (JSON string values only):
`NONE` | `VIOLATION_FAIRNESS` | `VIOLATION_LOYALTY` | `INTRODUCTION_HARM` | `OTHER`
Definition

`EthicalCatalystSecondary` captures additional ethical pressures present in the transition beyond the primary catalyst. It models ethical entanglement — cases in which multiple moral tensions co-occur and jointly intensify or complicate transformation.

Where:

- `ethical_catalyst_primary` identifies the dominant rupture driver,
- `EthicalCatalystSecondary` captures reinforcing or interacting violations.

This field does not redefine the primary rupture. It records structural moral complexity.

Structural Role
`EthicalCatalystSecondary` enables:

- modeling of compound ethical rupture
- analysis of moral density under perturbation
- comparison of single-axis vs multi-axis violations
- examination of how ethical stacking influences ΔVCB and ΔSAI

It preserves interpretive nuance without fragmenting the Event into multiple annotations.

Annotation Instructions

- Populate this field only when a second distinct ethical violation is clearly present.
- If the primary ethical catalyst fully explains the rupture, set: `"ethical_catalyst_secondary": "NONE"`.
- The secondary catalyst must represent a separate moral dimension, not a restatement of the primary.
- Use `OTHER` sparingly. When `OTHER` is used, record justification using the reserved key `_comment`.
- Do not invent new enum values. Ethical enums must be encoded as uppercase JSON strings. Abbreviations are not permitted.

Selection Rules
Use a secondary catalyst only when: the passage explicitly signals two distinct moral pressures; both materially influence the destabilization; and removing one would meaningfully alter the rupture interpretation.

Do not use this field for: emotional amplification alone; contextual detail that does not introduce independent ethical tension; or downstream consequences of the primary violation.

Example

A trusted colleague secretly manipulates performance data to secure a promotion.

- **Primary**: `"ethical_catalyst_primary": "VIOLATION_LOYALTY"`
- **Secondary**: `"ethical_catalyst_secondary": "VIOLATION_FAIRNESS"`
- **Justification**: The betrayal is relational (loyalty), but the systemic manipulation introduces procedural injustice (fairness).

3.7.3 Example Annotation

Text:

"When she realized her best friend had leaked the files to secure an undeserved promotion, her pulse quickened. She wanted to scream, but instead, she smiled tightly."

```
{
  "energy_gradient": 0,
  "ethical_catalyst_primary": "VIOLATION_LOYALTY",
  "ethical_catalyst_secondary": "VIOLATION_FAIRNESS"
}
```

Purpose of the Energetic & Ethical Dynamics Layer

This layer ensures that transformation is not treated as a neutral state change. It encodes why the rupture mattered and how the system mobilized energy in response. In **Measurement Mode**, `EnergyGradient` and `EthicalCatalyst` fields enable cross-corpus aggregation of rupture directionality and moral entanglement patterns under controlled calibration sets.

By separating affective force from ethical cause, the schema allows downstream systems to:

- Distinguish agitation from collapse
- Model moral pressure without collapsing into sentiment labels
- Track how ethical complexity shapes cognitive motion

Together with **Affective Energy**, **Liminality**, and **Time**, this layer completes the directional and causal anatomy of transformation.

3.7.4 Measurement Mode — EnergyGradient Calibration Set (EG-CS)

The Energetic & Ethical Dynamics layer functions both as a narrative instrument and as a measurement substrate. In narrative mode, `EnergyGradient` preserves interpretive nuance regarding how affective force mobilizes during rupture. In measurement mode, `EnergyGradient` must demonstrate inter-annotator stability and drift resistance under controlled conditions.

To support this requirement, the Unified Annotation Schema defines a frozen calibration artifact: **EnergyGradient Calibration Set (EG-CS)**.

Purpose

The EG-CS establishes anchor conditions for the directional encoding of affective force under perturbation. It serves to:

- Validate annotator alignment on `EnergyGradient` polarity
- Detect longitudinal drift in labeling behavior
- Ensure reproducibility across datasets and deployments
- Support downstream structural metrics (e.g., ΔPAD modeling, volatility normalization)

The EG-CS does not benchmark narrative quality. It benchmarks label stability.

Scope

- Applies only when **Measurement Mode** is active.
- Required for certification-level annotation workflows.
- Not required for exploratory or pedagogical use.

Composition Rules

The EG-CS consists of:

- A frozen set of short passages (2–4 sentences each).
- Exactly one dominant rupture per passage.
- Locked gold labels for:
  `energy_gradient`, `ethical_catalyst_primary`, `ethical_catalyst_secondary`.
- A brief structural rationale describing why the polarity is correct under schema rules.

The EG-CS is versioned (e.g., EG-CS v1.0) and must not be modified without incrementing the schema version.

Anchor Polarity Definitions (Normative)

Calibration items must exemplify:

- **`EXPANSIVE (1)`**: Outward affective discharge, escalation, mobilization, or expressive destabilization dominates the clause span.

- `CONTRACTIVE (-1)`: Inward containment, suppression, shutdown, dissociation, or collapse dominates the clause span.
- `OSCILLATING (0)`: Sustained alternation between expansion and contraction within the same clause span. Minimum requirement: two clearly observable polarity shifts.

Intensity alone does not determine polarity. Directionality governs classification.

Annotator Qualification Rule

When **Measurement Mode** is active:

- Annotators must complete EG-CS prior to production labeling.
- A minimum agreement threshold (implementation-retained) must be met.
- Recalibration may be administered periodically to detect drift.

Failure to meet threshold does not invalidate the schema. It suspends measurement-grade annotation status.

Audit & Drift Detection

EG-CS items may be: Reintroduced at fixed intervals; Embedded unobtrusively in annotation batches; or Used to compute stability metrics across annotators or time windows. Deviation from anchor labels signals: instructional ambiguity, concept drift, annotator fatigue, or taxonomy misinterpretation. Such deviations should trigger review, not silent correction.

Structural Status

EG-CS operates at the certification layer. It does not modify: The Core Triad, The Structural & Temporal Layer, The Liminality Layer, or the Energetic & Ethical field set. It validates their directional encoding integrity.

Location of Calibration Items

The frozen calibration passages and gold labels are published in:
Appendix A — EnergyGradient Calibration Set (EG-CS v1.0)
Appendix artifacts are normative in Measurement Mode and illustrative in Narrative Mode.

Design Principle

The inclusion of EG-CS reflects a core commitment of the **Unified Annotation Schema**: *Narrative richness must remain measurable without being reduced.*

Measurement Mode does not collapse meaning into scalar simplicity. It ensures that directional force is encoded consistently enough to support structural analysis of transformation under perturbation.

3.8 Subjective & Interpretive Layer — The Observer Context

All annotation is mediated by perception. Rather than treating interpretive variance as statistical noise, **Mentim** formalizes it as structured context.

The Subjective & Interpretive Layer encodes annotator orientation at the session level, enabling:

- systematic analysis of interpretive variance
- bias mapping across cohorts
- reproducibility auditing
- and longitudinal drift detection

This layer does **not** modify core annotation outputs. It contextualizes them. **Observer context becomes machine-legible metadata**.

3.8.1 Annotator Profile

AnnotatorProfile
Enum:
`LUMPER` | `SPLITTER` | `INTUITIVE` | `ANALYTICAL` | `GUARDIAN` | `REBEL`

Definition
`AnnotatorProfile` captures dominant cognitive orientation tendencies derived from a short, structured calibration survey administered prior to annotation. Profiles represent heuristic bias tendencies, not fixed identities. They remain stable for a defined annotation session unless re-calibrated.

Profiles are metadata. They do not influence label eligibility or schema interpretation rules.

Profile Descriptions

Profile	Orientation	Structural Tendency
`LUMPER`	Integrative	Collapses ambiguity into cohesive wholes; favors coherence preservation
`SPLITTER`	Differentiating	Prioritizes categorical precision; sensitive to boundary distinctions
`INTUITIVE`	Affective	Relies on felt coherence and gestalt signal detection
`ANALYTICAL`	Logical	Applies rule-based reasoning and explicit criteria
`GUARDIAN`	Stability-Oriented	Sensitive to fairness, norm violation, and structural preservation
`REBEL`	Disruption-Oriented	Sensitive to rupture, authenticity, and transformational instability

These descriptors define bias directionality, not competence.

Multi-Profile Rule

If an annotator scores $\geq$ 30% alignment with two profiles:

- Record both profiles in a JSON array (ordered by dominance).
- Flag the annotation session for optional reviewer analysis.
- Do not modify core labels solely due to profile mixture.

Example: `"annotator_profile": ["ANALYTICAL", "GUARDIAN"]`

Operational Status

`AnnotatorProfile` operates at the session metadata layer. It is optional in **Narrative Mode**, required for **Measurement Mode** cohorts, and recommended for cross-cohort comparative studies.

It must not: alter primary annotation fields, substitute for inter-rater agreement metrics, or function as a justification for inconsistent labeling.

Measurement Purpose

`AnnotatorProfile` enables: variance partitioning (content-driven vs orientation-driven disagreement), bias clustering analysis, calibration set sensitivity testing, drift detection across time, and interpretive topology modeling.

In **Mentim's** framework, disagreement is not automatically error. It may reflect structured interpretive orientation. By encoding observer stance explicitly, the schema preserves interpretive richness while maintaining audit traceability.

Design Principle

The Subjective & Interpretive Layer acknowledges a structural fact: *Meaning is co-constructed.* Measurement integrity does not require suppressing perspective. It requires making perspective legible.

The archetypal annotator profiles are surface representations of continuous orientation dimensions maintained within the **Mentim** calibration layer. These latent axes enable statistical modeling of interpretive variance without altering the public annotation grammar.

3.8.2 Interpretive Angle

InterpretiveAngle
Enum: `INTERNAL` | `EXTERNAL` | `OMNISCIENT` | `UNRELIABLE` | `EMPATHIC`

Instruction
Tag the dominant narrative vantage point expressed by the passage. This field encodes how the text positions knowing (who perceives, from what distance, with what reliability), not what is perceived.

Definitions

`INTERNAL`
First-person or close third-person perspective; access to private thought, sensation, or affect is foregrounded.
Signal: felt experience is primary evidence.

`EXTERNAL`
Observational stance; the passage reports actions, events, or outcomes with minimal interior access.
Signal: behavior and description dominate; interiority is inferred or absent.

`OMNISCIENT`
God's-eye narration; the text asserts broad knowledge beyond any single character's access.
Signal: sweeping certainty, global context, cross-character mind access.

`UNRELIABLE`
The perspective is distorted, contradictory, self-deceptive, or demonstrably misaligned with presented evidence.
Signal: internal inconsistency, motivated reinterpretation, unstable reality contact.

`EMPATHIC`
The passage foregrounds relational attunement and emotional resonance as an interpretive lens (care, mirroring, interpersonal sensing).
Signal: meaning is carried through felt relational contact rather than detached description.

Disambiguation Note (Normative)

`EMPATHIC` describes the text's stance. `INTUITIVE` (AnnotatorProfile) describes the annotator's orientation. Both may co-occur; do not treat them as redundant.

Selection Rule

Select exactly one value: the angle occupying $\geq 60\%$ of the clause span. If no angle reaches dominance, choose the angle that initiates the passage.

Purpose

By encoding interpretive vantage, **Mentim** enables:

- clustering disagreement by narrative framing (rather than treating it as error)
- analysis of framing effects on ethical and liminal judgments
- separation of intrinsic narrative ambiguity from annotator inconsistency

3.8.3 Example — Full Combined Record

Text:
“She smiled at him, pretending everything was fine, but the air between them had changed.”

```
{
  "state_1": {
    "archetype": "INNOCENT",
    "affective_energy": "STABLE_HIGH"
  },
  "event": {
    "event_type": "BETRAYAL",
    "ethical_catalyst_primary": "VIOLATION_OF_LOYALTY"
  },
  "state_2": {
    "archetype": "SEEKER",
    "affective_energy": "AGITATED_UNSTABLE"
  },
  "temporal_marker": "REVEAL_MOMENT",
  "time_signature": "SUSPENDED",
  "reality_interface": "RELATIONAL",
  "epistemic_boundary": "BREACHED",
  "perceptual_mode": "RELATIONAL",
  "energy_gradient": 1,
  "liminal_quality": ["AMBIGUOUS"],
  "liminal_flag": "IMPLICIT",
  "annotator_profile": "INTUITIVE",
  "interpretive_angle": "EMPATHIC"
}
```

Purpose of the Subjective & Interpretive Layer

This layer prevents artificial consensus. Rather than averaging disagreement into noise, **Mentim** records:

- who is perceiving the transition
- from which cognitive stance it is interpreted
- and how perspective shapes rupture classification

In doing so, interpretation becomes a modeled variable rather than an uncontrolled bias. **Mentim** does not erase subjectivity. It instruments it.

3.9 Metacognitive & Ontological Layer — The Certainty & Context

The Metacognitive & Ontological Layer captures how the annotator understands their own interpretation and situates each annotation within the broader ontological structure of the Unified Annotation Schema.

This layer explicitly encodes:

- meta-awareness (confidence calibration)
- structured sources of interpretive friction, and
- ontological positioning within the maturation spectrum

In **Mentim**, uncertainty is not noise — it is a first-class signal. This layer transforms human doubt into analyzable metadata.

3.9.1 Annotator Confidence

Confidence
Type: Float ∈ [0.0, 1.0]
Note: Round to the nearest 0.1. Arbitrary precision is not permitted.

Definition
The annotator's calibrated estimate of structural interpretive certainty for the full Narr-Atom. Confidence reflects clarity of rupture, boundary timing, archetypal shift, and ethical catalyst—not emotional intensity of the passage.

Instructions

- **1.0** — Complete structural clarity; no plausible competing interpretation.
- **0.5** — Multiple defensible interpretations exist.
- **0.0** — Forced or placeholder label due to insufficient context.

If `Confidence = 0.0`, populate `_comment` with the specific cause (e.g., "truncated excerpt," "missing pre-state," "ambiguous boundary timing").

Purpose

Provides a numeric trace of epistemic uncertainty that enables:

- disagreement clustering
- entropy estimation across corpora
- annotator calibration scoring
- weighting in downstream modeling
- longitudinal interpretive stability tracking

This converts subjective doubt into measurable variance.

3.9.2 AmbiguityReason

AmbiguityReason
Type: Free-text
Constraint: Maximum 280 characters. Plain text only.

Definition
A concise explanation of why confidence is reduced or why interpretation may be contested. `AmbiguityReason` does not restate the annotation. It explains the structural source of uncertainty.

Examples

- "Could be either BETRAYAL or DISCOVERY depending on unseen context."
- "Boundary timing unclear between REVEAL_MOMENT and LIMINAL_INTERVAL."
- "Tone suggests irony; narrator reliability uncertain."
- "Energy appears contractive but metaphorical language complicates interpretation."

Perceptual Dependency Guidance
`AmbiguityReason` should align with **PerceptualMode**:

- **RELATIONAL** → emphasize interpersonal tension
- **SENSORY** → emphasize embodied ambiguity
- **SYMBOLIC** → emphasize metaphorical uncertainty
- **ABSTRACT** → emphasize conceptual ambiguity

This enforces a core Mentim principle:
Perception shapes interpretation.

3.9.3 AmbiguityType

AmbiguityType
Enum:
`LINGUISTIC` | `ETHICAL` | `PERCEPTUAL` | `ARCHETYPAL` | `LIMINAL` | `TEMPORAL`

Definition
Specifies the dominant structural source of interpretive variance. `AmbiguityType` classifies where disagreement originates.

Instruction
Select `AmbiguityType` whenever: `Confidence < 1.0` OR disagreement is detected across

parallel annotations.

Dependency Rule — Disagreement Mapping
If disagreement occurs in specific fields, `AmbiguityType` must reflect the structural locus:

- Different `ethical_catalyst_primary` → **ETHICAL**
- Different `epistemic_boundary` → **LIMINAL**
- Different `reality_interface` → **PERCEPTUAL**
- Different `temporal_marker` or `time_signature` → **TEMPORAL**
- Different `archetype` assignments → **ARCHETYPAL**
- Unclear wording or phrasing ambiguity → **LINGUISTIC**

Priority Rule (Single Selection Required)
If multiple ambiguity types apply, select the one structurally closest to the Event node using this hierarchy:

ETHICAL → LIMINAL → TEMPORAL → PERCEPTUAL → ARCHETYPAL → LINGUISTIC

This prevents over-tagging and preserves measurement stability.

Purpose

Formalizes a central **UAS** principle: *Disagreement is structured data, not annotator error.* `AmbiguityType` operates alongside **PerceptualMode**:

- **AmbiguityType** → where uncertainty arises
- **PerceptualMode** → how it is experienced

Together, they transform interpretive variance into analyzable topology.

3.9.4 ArchetypeFamily

ArchetypeFamily
Type: Ordered string
Syntax: Use ASCII arrow → surrounded by spaces; no nested parentheses.

Example (PT-01 canonical value):
`"Child → Innocent → Seeker → Sage"`

Definition
`ArchetypeFamily` encodes the longitudinal developmental lineage within which the current archetype is situated. It anchors local State annotations to a broader maturation trajectory. This field situates individual transitions within structured developmental arcs, enabling cross-transition coherence analysis and longitudinal modeling of cognitive maturation.

PT-01 Constraint (Canonical Anchor)
For Proto-Transition 01 (PT-01), `ArchetypeFamily` is fixed and invariant.

Literal value (must not be altered, abbreviated, or reordered):
`"Child → Innocent → Seeker → Sage"`

PT-01 functions as the canonical rupture anchor of the ontology. Its lineage remains constant to preserve calibration stability across datasets and deployments.

Post-PT-01 Flexibility
For subsequent proto-transitions (PT-02+), `ArchetypeFamily` values must be selected from a version-controlled ontology registry. Free-form lineage creation is not permitted. Each registered family must specify:

- Origin state
- Canonical rupture pattern
- Expected maturation endpoint
- Stability norms for transition behavior

This preserves: cross-family comparability, developmental coherence, and measurement stability across deployments.

Purpose

`ArchetypeFamily` links individual State annotations to structured developmental trajectories, enabling:

- Cross-transition coherence modeling
- Archetypal maturation analysis
- Long-horizon stability tracking
- Alignment with psychological and ethical development frameworks

3.9.5 EthicalDomain

EthicalDomain
Enum:
`Fairness` | `Loyalty` | `Care` | `Authority`
Note: Capitalize first letter only; do not use uppercase.

Definition
`EthicalDomain` encodes the broader moral foundation informing the primary ethical catalyst identified in the Event node. It situates localized rupture within universal ethical dimensions, enabling cross-transition and cross-domain analysis.
Instructions

- Select one dominant domain per annotation.

- If multiple domains are implicated, select the domain most directly violated in the triggering Event.
- `EthicalDomain` must correspond to the declared `ethical_catalyst_primary`.

Purpose

`EthicalDomain` connects situational rupture to universal moral architectures, enabling:

- Cross–proto-transition comparison
- Ethical trend analysis across corpora
- Longitudinal moral drift modeling
- Integration with moral psychology research (e.g., Moral Foundations Theory)

This field abstracts the catalyst into a comparative moral axis without collapsing the Event into sentiment.

3.9.6 Example Annotation

Text:
"When she saw the results, she felt a rush of disbelief. Someone had cheated."

```
{
  "state_1": {
    "archetype": "INNOCENT",
    "affective_energy": "STABLE_HIGH"
  },
  "event": {
    "event_type": "DISCOVERY",
    "ethical_catalyst_primary": "VIOLATION_FAIRNESS"
  },
  "state_2": {
    "archetype": "SEEKER",
    "affective_energy": "AGITATED_UNSTABLE"
  },
  "energy_gradient": 1,
  "temporal_marker": "REVEAL_MOMENT",
  "liminal_flag": "EXPLICIT",
  "confidence": 0.8,
  "ambiguity_reason": "Narrator tone may imply resignation
                       rather than shock.",
  "ambiguity_type": "LINGUISTIC",
  "archetype_family": "Child -> Innocent -> Seeker -> Sage",
  "ethical_domain": "Fairness"
}
```

3.10 Administrative & Lineage Layer

(The "Provenance & Version Control")

The Administrative & Lineage Layer encodes the conditions under which an annotation was produced. Where prior layers describe transformation, force, temporality, and interpretation, this layer preserves continuity across schema evolution.

Structural accountability requires historical traceability. Annotations must remain interpretable not only within a narrative arc, but across:

- schema revisions
- proto-transition expansions
- annotator recalibration cycles
- longitudinal dataset growth

This layer ensures that annotations remain reproducible as the ontology matures.

3.10.1 Administrative Fields

UAS_Version
Type: String (Semantic Versioning: MAJOR.MINOR)
Example: `"1.1"`

Definition: Identifies the schema version used at the time of annotation.
Instruction:

- Use MAJOR.MINOR format.
- Patch levels are not recorded in Phase 1.
- Increment MAJOR when structural fields change.
- Increment MINOR when vocabularies expand without structural modification.

Purpose: Prevents interpretive drift when the codebook evolves.

ProtoTransition
Type: String
Example: `"PT-01"`

Definition: Indicates the narrative transition class under which the annotation is scoped.
Instruction:

- Must match a defined proto-transition in the current schema version.
- Required field.
- Fixed to `"PT-01"` for Phase 1 dataset.

Purpose: Ensures prototype-scoped comparability and prevents cross-transition collapse in analysis.

Annotator_ID (Optional)
Type: String (16-character truncated SHA-256 hash)
Definition: Pseudonymous identifier enabling rater-lineage analysis without exposing identity.
Constraints:

- No personally identifiable information.
- Derived from salted internal identifier.
- Stable across batches unless recalibrated.

Purpose: Enables bias tracking, calibration drift detection, profile-based disagreement clustering, and longitudinal rater stability analysis.

Timestamp (Recommended)
Type: ISO 8601 string

Example: `"2026-02-16T15:42:00Z"`

Definition: Records the moment of annotation.
Purpose: Supports revision audits, inter-version comparison, temporal clustering of interpretive shifts, and controlled calibration studies.

Ontological Rationale

This layer is not clerical. It preserves structural continuity across schema evolution. Without provenance, interpretive systems drift silently. Without versioning, calibration collapses. Without lineage, disagreement cannot be contextualized.

The **Administrative & Lineage Layer** ensures that:

- meaning remains historically anchored
- ontology expansion remains backward-compatible
- annotation datasets remain governance-ready

Structural accountability requires that the measurement instrument itself be accountable. This layer encodes that accountability.

3.11 Field Reference Table

This table defines the complete annotation field set for Proto-Transition 01 (PT-01). It functions simultaneously as:

- A normative data dictionary (field name + allowable values)
- A conceptual mapping from theoretical dimensions (Time, Affective Energy, Liminality, Archetype, Ethics) to operational fields

All enums are case-sensitive and must match exactly as specified.

Developmental Adequacy (D-SES) Clarification

UAS v1.0 defines the full required field set for PT-01. D-SES (Narr-Atom-D) introduces:

- No additional required fields
- No modification to the SES grammar
- No alteration of the data dictionary

Any D-SES designation operates strictly at the certification layer. Future versions may introduce optional overlay fields for expectation formation, salience differentiation, and confidence modulation without modifying the PT-01 core field set.

I. Core Triad (The "What")

`state_1.archetype`
Description: Archetypal configuration prior to rupture.
Allowed Values: `INNOCENT`.

`state_1.affective_energy`
Description: Baseline energetic equilibrium.
Allowed Values: `STABLE_HIGH`.

`event.type`
Description: Nature of epistemic rupture.
Allowed Values:

- `BETRAYAL`
- `DISCOVERY`

`event.ethical_catalyst_primary`
Description: Primary ethical driver.
Allowed Values:

- `VIOLATION_FAIRNESS`
- `VIOLATION_LOYALTY`
- `INTRODUCTION_HARM`

`state_2.archetype`
Description: Archetypal configuration post-rupture.
Allowed Values: `SEEKER`.

`state_2.affective_energy`
Description: Disequilibrium and search orientation.
Allowed Values: `AGITATED_UNSTABLE`.

II. Structural & Temporal Layer (The "When & How")

`temporal_marker`
Description: Narrative position relative to rupture.
Allowed Values:

- `PRE_EVENT`
- `REVEAL_MOMENT`
- `LIMINAL_INTERVAL`
- `POST_REVEAL_AWARENESS`

`time_signature`
Description: Subjective temporal texture.

Allowed Values:

- `STAGNANT`
- `FLOWING`
- `CYCLICAL`
- `ACCELERATED`
- `SUSPENDED`

`reality_interface`
Description: Origin of rupture.
Allowed Values:

- `INTERNAL`
- `EXTERNAL`
- `PERCEIVED`
- `ABSTRACT`

`epistemic_boundary`
Description: Status of worldview.
Allowed Values:

- `TRUSTED`
- `BREACHED`
- `FUGITIVE`
- `REFORMED*`

**REFORMED is not permitted within PT-01.*

`perceptual_mode`
Description: Dominant perceptual channel.
Allowed Values:

- `SENSORY`
- `IMAGINED`
- `SYMBOLIC`
- `ABSTRACT`
- `RELATIONAL`

III. Liminality Layer (The "Between")

`liminal_quality`
Description: Tone of transition (array permitted).
Allowed Values:

- `SUDDEN`
- `GRADUAL`

- `AMBIGUOUS`
- `PAINFUL`
- `JOYOUS`
- `NUMB`
- `TRANSFORMATIVE`

`liminal_flag`
Description: Visibility of liminal signal.
Allowed Values:

- `EXPLICIT`
- `IMPLICIT`
- `ABSENT`

IV. Energetic & Ethical Dynamics (The "Force & Why")

`energy_gradient`
Description: Directional affective movement.
Allowed Values: `1 | -1 | 0`.
Values for energy_gradient must be integers, not strings.

`ethical_catalyst_secondary` (optional)
Description: Secondary ethical pressure.
Allowed Values:

- `NONE`
- `VIOLATION_LOYALTY`
- `VIOLATION_FAIRNESS`
- `INTRODUCTION_HARM`

V. Subjective & Interpretive Layer (The "Who & From Where")

`annotator_profile`
Description: Annotator orientation.
Allowed Values:

- `LUMPER`
- `SPLITTER`
- `INTUITIVE`
- `ANALYTICAL`
- `GUARDIAN`
- `REBEL`

`interpretive_angle`

Description: Narrative vantage point.

Allowed Values:

- `INTERNAL`
- `EXTERNAL`
- `OMNISCIENT`
- `UNRELIABLE`
- `EMPATHIC`

VI. Metacognitive & Ontological Layer (The “Certainty & Context”)

`confidence`

Description: Self-assessed certainty (round to 0.1).

Allowed Values: `Float [0.0-1.0]`.

`ambiguity_reason`

Description: Qualitative explanation.

Allowed Values: `Free text (≤280 chars)`.

`ambiguity_type`

Description: Dominant variance source.

Allowed Values:

- `LINGUISTIC`
- `ETHICAL`
- `PERCEPTUAL`
- `ARCHETYPAL`
- `LIMINAL`
- `TEMPORAL`

`archetype_family`

Description: Developmental lineage (fixed in PT-01).

Allowed Values: `"Child → Innocent → Seeker → Sage"`.

`ethical_domain`

Description: Broader moral domain.

Allowed Values: `Fairness | Loyalty | Care | Authority`.

VII. Administrative & Lineage Layer

`UAS_Version`

Description: Schema version (e.g., "1.0").

Type: `String`.

`ProtoTransition`
Description: Indicates the narrative transition class.
Value: `"PT-01"`.

`Annotator_ID` (optional)
Description: Pseudonymous lineage ID.
Type: `16-char hash`.

`Timestamp` (optional)
Description: Annotation time.
Type: `ISO 8601`.

`phenomenological_density (Φ)`
Description: Experiential richness indicator.
Type: `Float [0-1]`.

PhenomenologicalDensity (Φ)

Type: Float $\in$ [0.0, 1.0].
Note: Round to two decimal places; no arbitrary precision.

Definition

`PhenomenologicalDensity (Φ)` estimates the degree of experiential channel activation present within a single Narr-Atom. It captures how many distinct experiential dimensions are actively engaged during the transition, without evaluating their correctness, intensity, or normative adequacy.

Computation Rule

Φ is calculated as: **(Number of active channels) ÷ 4**. Where the counted channels are:

- `EnergyGradient` (directional affective movement)
- `LiminalQuality` (qualitative tone of crossing)
- `EthicalCatalystPrimary` (moral driver of rupture)
- `PerceptualMode` (dominant awareness channel)

A channel is considered active if it is explicitly encoded in the annotation record and not `NULL`.
Example: All four channels present → $\Phi = 1.00$; Three channels present → $\Phi = 0.75$; Two channels present → $\Phi = 0.50$.

Interpretive Scope

Φ is: **Descriptive**, **Corpus-analytic**, and **Structural**.

Φ is **not**:
A quality score, a maturity index, a confidence modifier, or a weighting variable in SAI or downstream evaluation.

Constraint
`PhenomenologicalDensity (Φ)` must never automatically influence:

- Structural Accountability Index (SAI)
- Developmental Adequacy (D-SES) designation
- Confidence weighting
- Annotation validity

Its function is observational, not normative.

Purpose

Φ enables:

- Cross-corpus comparison of experiential richness
- Detection of sparsely encoded vs densely encoded transitions
- Analysis of how narrative segments distribute experiential channels

It does not collapse interpretive richness into scalar judgment. In alignment with **Mentim** principles, Φ preserves descriptive plurality without imposing evaluative hierarchy.

In alignment with **Mentim** principles, Φ preserves descriptive plurality without imposing evaluative hierarchy.

4. Annotation Guidelines & Inter-Rater Reliability (IRR)

IRR Scope Note
Inter-rater reliability (IRR) procedures defined in this section apply only to core SES annotations specified in UAS v1.1 (PT-01).

Developmental adequacy designations (D-SES / Narr-Atom-D) operate at the certification layer and are explicitly excluded from IRR requirements in this version. **Core SES fields are the sole basis for IRR scoring in Phase 1**.

4.1 Document-Level Classification (PT-01)

Before span-level annotation, annotators must classify each passage at the document or excerpt level. This classification determines whether the passage plausibly instantiates Proto-Transition 01 (PT-01: Ethical Coherence Rupture / The Shattering of Innocence) and governs downstream ambiguity handling.

Each passage must be assigned exactly one of the following categories:

CLEAR_POSITIVE

Definition: An unambiguous instance of the Innocent → Reveal → Seeker triad, where all three components are explicitly present or strongly implied and occur in coherent temporal order.

Required Anchors:

- Baseline assumed coherence or trust (Innocent)
- Concrete ethical rupture (BETRAYAL, DISCOVERY, or INTRODUCTION_OF_HARM)
- Active reorientation toward inquiry (Seeker)

Diagnostic Test: If the Reveal were removed, would the subject's worldview remain structurally intact?

- If removal eliminates transformation → CLEAR_POSITIVE
- If removal changes little → reassess classification

CLEAR_NEGATIVE

Definition: A negative event or emotional disturbance that does not rupture foundational coherence and does not initiate active epistemic reorientation.

Typical Features:

- Complaint or frustration without ethical violation
- Routine setback
- Emotional response without worldview destabilization
- Informational update without structural reorganization

Exclusion Rule: If no ethical rupture is present, the passage cannot qualify as PT-01.

AMBIGUOUS

Definition: A boundary case where the presence, strength, or structural completeness of the PT-01 triad is contestable.

Common Triggers:

- Partial triad (e.g., Reveal present; Seeker state unclear)
- Competing plausible archetype interpretations
- Tone obscures intent (irony, unreliable narration)
- Ethical rupture implied but under-specified

Annotator Action:

- Apply span-level annotations as usual
- Set `liminal_flag` = `IMPLICIT` when appropriate
- Populate `ambiguity_reason`
- Assign `ambiguity_type` (mandatory if confidence < 1.0)

- Use OTHER only after exhausting enum options

Note: Ambiguity is treated as structured data, not annotator error.

INVERSE_TRANSITION

Definition: A regression from Seeker back toward Innocent through denial, minimization, repression, or rationalized restoration of prior coherence.

Typical Signals:

- Explicit dismissal of rupture
- Reinstatement of prior assumptions without integration
- Statements such as: "It was nothing."; "I shouldn't have questioned it."; "Everything is fine."

Constraint: Do not assign INVERSE_TRANSITION unless a prior CLEAR_POSITIVE PT-01 instance exists for the same subject or narrative episode. This category captures defensive restoration of innocence, not resolution.

Quick Decision Heuristics (IRR Aids)

Annotators may use the following decision flow:

1. Is there an explicit or strongly implied ethical rupture?
   - If no → **CLEAR_NEGATIVE**
2. If yes, does the subject actively seek reorientation, inquiry, or reinterpretation?
   - If yes → **CLEAR_POSITIVE**
3. If evidence is partial, unstable, or interpretation-dependent → **AMBIGUOUS**
4. If the text depicts rollback after prior seeking → **INVERSE_TRANSITION**

IRR Protocol Note

Document-level classification is used to: normalize span-level disagreement, improve inter-annotator calibration, and distinguish structural absence from interpretive instability. Disagreements at this stage must be logged and reviewed. They must not be resolved through majority vote without structured discussion.

4.2 Inter-Rater Reliability (IRR) — Structural Interpretation Model

In conventional annotation systems, inter-rater reliability (IRR) is treated as a proxy for annotation correctness. High agreement is equated with quality; disagreement is treated as error, noise, or annotator failure.

The Unified Annotation Schema adopts a different position:

- Within Mentim, IRR is not interpreted as a measure of correctness.

- It is interpreted as a measure of structural stability versus interpretive complexity within the annotated material.
- Agreement reflects convergence of perception under a defined ontology.
- Divergence reflects ethical tension, archetypal ambiguity, liminal instability, or unresolved structural complexity in the source text.

IRR therefore functions as a diagnostic instrument about the structure of meaning itself — not merely about annotator performance.

IRR Interpretation Bands

The following qualitative bands apply to standard agreement statistics (e.g., Krippendorff's α, Cohen's κ). These bands are interpretive, not punitive.

α or $\kappa \geq 0.80$ — Convergence Zone
Interpretation: Strong structural stability across annotators.
Characteristics: High cross-rater consensus, minimal ambiguity, clear ethical/archetypal signaling, low liminal volatility.
Analytical Role: Suitable for calibration baselines, useful for training controlled classifiers, limited discovery potential.
Example: A direct betrayal with explicit trust, explicit violation, and explicit reorientation.

α or $\kappa = 0.60$–0.79 — Resonance Band
Interpretation: Stable core agreement with bounded interpretive variation.
Characteristics: Agreement on primary Event and Archetype, divergence in secondary dimensions (energy gradient, liminality tone, interpretive stance).
Analytical Role: Ideal modeling material, preserves nuance without structural instability, indicates shared moral grammar with stylistic variance.
Example: Annotators agree on fairness violation but differ on whether affect is expansive or oscillating.

α or $\kappa = 0.40$–0.59 — Fertile Ground
Interpretation: Significant interpretive divergence under a shared schema. **This is Mentim's primary discovery zone**. **Characteristics**: Competing plausible archetypes, ethical ambiguity/entanglement, liminal instability.
Analytical Role: Signals ontological richness, guides schema refinement, identifies material suitable for introspective modeling.
Example: Disagreement over whether confusion represents active seeking (Seeker) or defensive rollback (Inverse Transition).

α or $\kappa < 0.40$ — Paradox Zone
Interpretation: Extreme divergence requiring investigation.

Possible Explanations: Genuinely multivalent text, conflicting ethical domains, archetypal oscillation.
Analytical Response: Conduct structured review, examine `AmbiguityType` clustering, evaluate ontology extension.
Example: A passage plausibly interpretable as both liberation and betrayal depending on stance.

Guiding Principle

Mentim does not optimize for unanimity; it optimizes for traceable structure under plurality.

- High IRR identifies stabilized, culturally shared meaning.
- Moderate IRR identifies nuance and stylistic inflection.
- Low IRR identifies tension, transformation, and ontological frontier.

Inter-rater disagreement is therefore preserved as structured signal — not discarded as defect.

4.3 Complexity-Aware Metrics — Disagreement as Structured Signal

Inter-rater reliability (IRR) summarizes overall convergence. Complexity-aware metrics characterize the internal structure of disagreement.

Within the Unified Annotation Schema, disagreement is not reduced to a single coefficient. It is decomposed into measurable components that reveal whether interpretive variance is diffuse, polarized, clustered, or dimension-specific.

For each annotated item — and for each eligible dimension — the following metrics are computed:

Disagreement Index ($\mathcal{D}$)

Definition: $\mathcal{D}$ is a normalized measure of interpretive dispersion across annotators. It captures how widely interpretations are distributed across the allowable label space.
Operationalization (by data type):

- **Categorical fields**: Normalized entropy (base-2). Entropy is divided by maximum possible entropy for the category size to scale $\mathcal{D} \in [0, 1]$.
- **Ordinal fields**: Coefficient of variation (σ/μ), normalized to [0,1] under bounded scale assumptions.
- **Multi-label fields**: Mean Bernoulli variance across labels.

Interpretive Meaning: $\mathcal{D}$ measures breadth. Low $\mathcal{D} \rightarrow$ strong convergence; Moderate $\mathcal{D} \rightarrow$ bounded diversity; High $\mathcal{D} \rightarrow$ multiple plausible readings distributed across categories. **High $\mathcal{D}$ does not imply annotator failure; it indicates interpretive dispersion**.

Polarization (P)

Definition: P measures whether disagreement is diffuse or structured into opposing camps. Where $\mathcal{D}$ captures how many interpretations exist, P captures whether disagreement forms clusters.

Operationalization (by data type):

- **Categorical fields**: Top-two class proportion split.
- **Ordinal fields**: Bimodality coefficient or two-cluster separation index.
- **Multi-label fields**: 90th-percentile pairwise Jaccard distance across annotators.

Interpretive Meaning: Low $P \rightarrow$ disagreement scattered around a shared center; High $P \rightarrow$ disagreement structured into competing interpretive poles. Polarization is especially informative in ethical, archetypal, and liminal dimensions.

Complexity Flagging (Default Heuristic)

Initial screening rule (domain-tunable; threshold deviations ±0.10 must be documented in `_comment`):

Flag as High-Complexity Passage if:

- $\mathcal{D} > 0.55$
- $P > 0.60$

Flagged passages are routed to: adjudication review, ontology refinement, qualitative interpretive audit, or governance modeling datasets. These cases frequently correspond to:

- ethical entanglement
- archetypal ambiguity
- liminal instability
- epistemic fracture under perturbation

Structural Interpretation of Metrics

Metric	What It Measures	What It Reveals
IRR (α/κ)	Overall convergence	Structural stability
$\mathcal{D}$	Dispersion	Interpretive breadth
P	Cluster formation	Camp division
$\mathcal{D} + P$ high	Structured conflict	Ontological tension

Ambiguity–Disagreement Integration

Quantitative disagreement metrics are explicitly linked to qualitative annotation. Each record includes an `AmbiguityType`, which classifies the dominant source of interpretive variance: **ETHICAL, LIMINAL, TEMPORAL, PERCEPTUAL, ARCHETYPAL, LINGUISTIC**.

When computing $\mathcal{D}$ and P across a corpus, results are grouped by `AmbiguityType`. This allows the system to detect not only how much disagreement exists, but also where it concentrates.

Example patterns:

- High $\mathcal{D}$ in ETHICAL → moral domain instability.
- High P in ARCHETYPAL → identity polarization.
- High $\mathcal{D}$ + high P in LIMINAL → unstable transition boundary.

This converts disagreement into structured epistemic telemetry.

Paradigm Statement

Conventional annotation systems minimize disagreement. **Mentim models disagreement**. In this framework:

- Convergence indicates stabilized meaning.
- Dispersion indicates plural plausibility.
- Polarization indicates structured ethical or ontological tension.

Uncertainty is not removed; it is mapped. Disagreement becomes a measurable feature of cognitive structure under perturbation.

4.3.1 Training for Interpretive Range and Modeling Disagreement

The Unified Annotation Schema (UAS) reconceptualizes Inter-Rater Reliability (IRR) as a map of interpretive space rather than a proxy for correctness.

In this framework, low agreement is not inherently a defect. It is frequently the structural signature of conceptual richness, ethical entanglement, or liminal transition. Accordingly, both annotation training and downstream model training are explicitly designed to preserve and learn from ambiguity rather than suppress it.

A. Annotation Training — Structured Ambiguity Exposure

Annotator training under UAS intentionally incorporates passages that support multiple defensible interpretations, particularly along the Ethical, Archetypal, and Liminal dimensions. These include:

- events plausibly constituting both a Violation of Fairness and a Violation of Loyalty
- characters oscillating between Innocence and performative or defensive ignorance
- transitions in which liminality is experientially present but narratively implicit
- passages where tone destabilizes archetypal assignment

Annotators are instructed to:

- select the most defensible primary label
- record secondary ethical or interpretive influences where applicable
- assign `Confidence` values reflecting genuine epistemic certainty

- document interpretive friction in the `AmbiguityReason` field
- assign `AmbiguityType` when divergence is plausible

The resulting ensemble of labels, confidence values, and qualitative explanations constitutes the item's **epistemic signature** — a structured record of how human interpretation distributes across a shared narrative stimulus. **This signature is preserved, not collapsed**.

B. Modeling Disagreement — Distributional Fidelity
UAS explicitly rejects the assumption that intelligence consists in producing a single canonical label. Instead, downstream systems are trained to model the distribution of human interpretations associated with each item. Under this paradigm:

- training objectives reward fidelity to observed variance
- disagreement is treated as structured entropy rather than annotation noise
- high-variance regions are interpreted as indicators of liminality, ethical complexity, or ontological instability

Performance is assessed not solely by agreement with a majority label, but by the system's ability to approximate:

- the dispersion (breadth) of interpretation
- the polarization (clustering) of interpretation
- the qualitative sources of disagreement (`AmbiguityType` alignment)

The objective is not consensus reproduction; it is distributional alignment.

C. Epistemic Implication
Within the UAS framework, intelligence is not defined as certainty reproduction. It is defined as structural fidelity to the human spectrum of interpretation. A system trained under UAS is expected to:

- recognize where meaning fractures
- represent ambiguity without premature collapse
- model ethical and archetypal tension explicitly
- preserve structural continuity while operating under interpretive uncertainty

This approach directly advances **Mentim's governing principle**: *Cognition is accountable not when it avoids contradiction, but when it remains coherent while modeling it.*

4.4 Adjudication & Revision Loop

The Unified Annotation Schema treats adjudication not as error correction, but as ontology maintenance.

This section defines the structured process by which disagreement is:

- interpreted,
- preserved, or
- used to refine the schema itself.

Adjudication exists to distinguish interpretive richness from schema insufficiency.

4.4.1 Review Cadence

Frequency: After every 50 annotated items
Trigger Condition: Completion of a batch with computed $\mathcal{D}$ and P metrics

This cadence ensures:

- statistical stability (patterns over outliers),
- operational responsiveness,
- controlled schema evolution.

Batches smaller than 30 items SHALL NOT trigger formal adjudication unless flagged by manual override.

4.4.2 Triage and Prioritization

Items **SHALL** be ranked by:

- Disagreement Index ($\mathcal{D}$)
- Polarization (P)

Primary review set:
Top 10–20 items with highest combined $\mathcal{D}$ and P values. These items most frequently correspond to:

- liminal transitions,
- ethical entanglement,
- archetypal ambiguity,
- schema boundary exposure.

Lower-variance items may be sampled randomly for control comparison.

4.4.3 Group Review Protocol

For each prioritized item, reviewers **SHALL** examine:

- Full annotation records
- `AmbiguityReason` entries
- `AmbiguityType` assignments

- Confidence distributions
- Divergence across Ethical, Archetypal, Liminal, and Temporal dimensions
- Associated $\mathcal{D}$ and P metrics

The objective is not forced convergence. The guiding diagnostic question is: *What type of disagreement is present, and what does it reveal about the ontology?*

4.4.4 Adjudication Outcomes

Each reviewed item **MUST** result in one of the following determinations:

1. True Multivalence (Preserve Variance)
Criteria:

- Disagreement reflects multiple legitimate interpretive positions.
- No definitional inconsistency or cue ambiguity is identified.

Action:

- All labels are preserved.
- The item (or distilled variant) **SHALL** be added to the codebook as a "Both-Valid" reference case.
- Confidence and `AmbiguityType` patterns are retained.

Note: These cases become calibration anchors for interpretive plurality.

2. Definition or Cue Gap (Refine Schema)
Criteria: Divergence arises from unclear definitions, missing cues, or overlapping category boundaries.
Permissible Actions:

- Refine wording of existing dimensions
- Add positive or exclusionary cues
- Introduce clarifying examples
- Tighten boundary rules

Constraint:

- Revisions **SHALL** be minimal and scoped narrowly.
- No retroactive reclassification of previously valid interpretations is permitted.

4.4.5 Versioning and Recalibration

All schema changes **SHALL** trigger:

UAS_Version increment:

- **Minor version**: clarifications or examples
- **Major version**: new dimensions or breaking redefinitions

Changelog entry documenting:

- What changed
- Why it changed
- Which ambiguity patterns motivated revision

After any version increment:

- Annotators **SHALL** complete recalibration using 10 anchor items.
- Anchor items **SHOULD** be drawn from True Multivalence cases where possible.
- Prior annotations remain valid under their original `UAS_Version`.

Historical integrity MUST be preserved.

4.4.6 Governance Principle

The adjudication loop operates under one non-negotiable rule:
Disagreement is resolved only when it reflects schema failure — not when it reflects human meaning.

Interpretive Logic:

- If interpretive variance reflects legitimate plurality, it is preserved.
- If variance reflects structural ambiguity in the ontology, the ontology evolves.

The Unified Annotation Schema is therefore a living system:

- **stable in grammar**,
- **adaptive in refinement**,
- **accountable in revision**.

4.5 Rater Protocol

This protocol governs how human judgment is sampled, contextualized, and preserved within the Unified Annotation Schema.

The objective is not consensus. The objective is faithful capture of interpretive distributions.

Raters function as structured sensors of meaning, not arbiters of correctness.

4.5.1 Rater Density

Target raters per item: 6
Permitted range: 5–7

This range is calibrated to:

- preserve minority readings,
- reduce majority erasure effects,
- enable stable estimation of $\mathcal{D}$ and P,
- maintain operational feasibility.

Items annotated by fewer than 5 raters SHALL NOT be included in disagreement metric computation.

4.5.2 Annotation Overlap

Minimum overlap requirement:
$\geq 25\%$ of total items **MUST** be double- or triple-annotated.

Overlap serves four purposes:

- Compute IRR (α / κ)
- Compute Disagreement Index ($\mathcal{D}$)
- Compute Polarization (P)
- Detect cohort-based interpretive drift

Overlap is treated as a diagnostic instrument, not a quality penalty.

Low agreement within overlap does not invalidate the item. It triggers interpretive analysis under §4.3 and §4.4.

4.5.3 Annotator Context Capture

At the beginning of each annotation session, raters **MUST** record:
`AnnotatorProfile`

- (e.g., LUMPER, SPLITTER, INTUITIVE, ANALYTICAL, GUARDIAN, REBEL)

`InterpretiveAngle`

- (e.g., INTERNAL, EXTERNAL, EMPATHIC, OMNISCIENT, UNRELIABLE)

Profiles persist across sessions unless recalibration explicitly reassigns them.

These fields serve to:

- distinguish perspective-driven variance from schema ambiguity,
- cluster disagreement by cognitive style,
- model orientation-dependent interpretation.

(See §3.8 Subjective & Interpretive Layer.)

4.5.4 Confidence and Ambiguity Requirements

For items classified as **AMBIGUOUS** or **INVERSE_TRANSITION**, raters **MUST** provide:

Primary Metrics:

- `Confidence` (float $\in$ [0.0–1.0], rounded to nearest 0.1)
- `AmbiguityReason` (free-text, $\leq$ 280 characters)

Constraints:

- `Confidence` $= 0.0$ is permitted only when accompanied by mandatory `AmbiguityReason`.
- `AmbiguityType` **MUST** be assigned when `Confidence` < 1.0.

This ensures that uncertainty is:

- explicit rather than implicit,
- attributable to structured interpretive factors,
- analyzable rather than discarded.

Silence is not allowed where ambiguity is perceived.

4.5.5 Cohort Integrity

If annotator pools are refreshed or expanded:

- A minimum 10-item recalibration set **MUST** be administered.
- $\mathcal{D}$ and P distributions **SHALL** be compared against prior cohort baselines.
- Significant distribution shifts ($> \pm 0.10$ in mean $\mathcal{D}$ across core fields) **MUST** be logged.

This protects longitudinal interpretive stability.

4.5.6 Guiding Principle

The rater is not a judge of correctness. **The rater is a calibrated sensor of transformation**.

The Rater Protocol ensures that annotations reflect how humans actually perceive rupture, liminality, and reorganization — preserving the epistemic richness required for structural accountability modeling.

In Mentim:

- **Disagreement is preserved**.

- **Uncertainty is formalized**.
- **Meaning is sampled — not flattened**.

4.6 Reporting

Each annotation batch **MUST** be accompanied by a structured reporting package summarizing:

- data quality,
- interpretive diversity,
- disagreement structure,
- rater behavior.

Reporting exists to characterize epistemic texture — not to enforce consensus.

Reports are suitable for:

- research publication,
- model training diagnostics,
- governance and audit review.

4.6.1 Document-Level Reporting

For each completed batch, the following **MUST** be reported:

1. Class Distribution
Proportion and raw count for:

- `CLEAR_POSITIVE`
- `CLEAR_NEGATIVE`
- `AMBIGUOUS`
- `INVERSE_TRANSITION`

Report both percentages and absolute counts.
Purpose: Establishes the structural composition of the corpus and prevents silent skew toward trivial or stable cases.

2. Inter-Rater Reliability (IRR)
Report:

- Global α or κ (rounded to two decimal places)
- 95% confidence intervals when $n \geq 30$
- IRR by classification class
- IRR by core dimension (where applicable)

Note: IRR is descriptive, not normative.
Purpose: Characterizes convergence zones versus high-interpretation zones.

3. Complexity Metrics

Report:

- Mean and median Disagreement Index ($\mathcal{D}$)
- Mean and median Polarization (P)
- Percentage of high-complexity items ($\mathcal{D} > 0.55$ AND $P > 0.60$)

If threshold adjustments are made (±0.10), document justification in `_comment`.

Purpose: Reveals whether the dataset is: **stability-dominant, liminal-heavy, ethically entangled, or polarization-prone**.

Note: This is a measure of interpretive density, not dataset "quality".

4.6.2 Span-Level, Dimension-Specific Reporting

For each major dimension and layer:

- `Archetype`
- `Affective Energy`
- `Ethical Catalyst`
- `Time (TemporalMarker + TimeSignature)`
- `Liminality`
- `RealityInterface`
- `EpistemicBoundary`

Report:

- Agreement metric (α, κ, or appropriate statistic)
- Mean and median $\mathcal{D}$
- Mean and median P
- Distribution across `AmbiguityType`
- `NULL` counts (reported explicitly; never imputed)

Purpose: Identifies where interpretive variance concentrates:

- Ethical disagreement
- Liminal instability
- Archetypal ambiguity
- Temporal confusion
- Perceptual divergence

This guides ontology refinement and model calibration.

4.6.3 Rater Diagnostics

Each reporting package **SHOULD** include rater-level analytics.

1. Confusion Matrices

- By classification class
- By core dimension

2. Drift Analysis

- Within-batch drift
- Cross-batch drift
- Rolling $\mathcal{D}$ and P averages

Flag sustained directional shifts ($> \pm 0.10$ in mean $\mathcal{D}$ across batches).

3. Correlation Analysis
Report correlations between:

- `AnnotatorProfile` and label choice
- `InterpretiveAngle` and `AmbiguityType`
- Confidence and Disagreement contribution
- `EnergyGradient` and `LiminalQuality` variance

Purpose: Distinguishes **structured epistemic ambiguity** from **rater fatigue, bias, or calibration drift**.

Note: Calibration interventions occur only when drift is systematic, not when disagreement reflects meaningful multivalence.

4.6.4 Reporting Principle

Reporting does not judge correctness. **Reporting maps the structure of disagreement**. All metrics must be interpreted as:

- signals of meaning density,
- indicators of liminal pressure,
- evidence of ontological complexity,
- reflections of cognitive plurality.

The goal is not to reduce disagreement. The goal is to understand it.

4.7 Quality Control Without Convergence

Purpose: The Unified Annotation Schema explicitly rejects convergence as the sole indicator of data quality. In domains involving ethics, narrative identity, liminality, and affect, forced agreement collapses structure rather than validating it.

Accordingly, Mentim defines quality not as unanimity of labels, but as:

- stability of interpretive distributions
- intelligibility of disagreement
- reproducibility of variance

Principle: Structure Over Consensus
An annotation batch is considered high quality when:

- Interpretive variance is bounded and interpretable
- Disagreement patterns are consistent across annotator cohorts
- Divergence aligns with declared `AmbiguityType` and `PerceptualMode`
- High-complexity items remain high-complexity under re-annotation

Note: Data quality is degraded when disagreement is erratic, unpatterned, or uncoupled from the ontology.

Quality Control Criteria (Non-Convergent)

Quality is evaluated using structural controls — not consensus thresholds.

1. Minimum Annotation Density: Each item must be annotated by ≥ 5 raters (Target: 6).
Rationale: Distributions must be observable. Point estimates are insufficient.

2. Disagreement Coherence: Items with elevated $\mathcal{D}$ and P must exhibit stable `AmbiguityType` assignment and show consistent domains of disagreement (e.g., **ETHICAL** vs **PERCEPTUAL**). Shifting ambiguity classifications indicate annotation failure, not conceptual richness.

3. Reproducibility of Variance: High-complexity items are reintroduced in later batches. Quality is confirmed when $\mathcal{D}$ and P remain within ±10–15% relative to prior batch mean, and dominant `AmbiguityType` remains stable.

4. Annotator Profile Balance: If one `AnnotatorProfile` exceeds 60% of the cohort, rebalancing is required. Quality requires plural cognitive perspectives; homogeneity reduces interpretive depth.

5. Confidence–Variance Alignment: Low `Confidence` scores should correlate with elevated $\mathcal{D}$ and P, and explicit `AmbiguityReason` entries. **Confidence is a meta-signal and must track variance**.

Explicit Failure Modes

The following constitute quality failure — even when IRR is high:

- Artificial convergence due to over-restrictive definitions
- Suppression of ambiguity through adjudication pressure
- Disagreement unlinked to declared `AmbiguityType`

- Annotator drift unexplained by profile shifts or retraining
- Cohort homogeneity masking interpretive diversity

Output Classification

Each batch receives one of the following status designations:

- **Structurally Sound**: Disagreement is patterned, interpretable, and reproducible.
- **Structurally Ambiguous (Acceptable)**: High variance present; suitable for ambiguity modeling.
- **Structurally Unstable (Hold)**: Variance is unpatterned or incoherent. Escalate to adjudication (§4.4) within 48 hours.

Governance Statement: Mentim's quality control framework preserves complexity without sacrificing rigor. This framework enables interpretive intelligence, not merely predictive accuracy.

4.8 Dataset Readiness, Release, and Governance Controls

Purpose: This section defines the conditions under which a Unified Annotation Schema (UAS) dataset is:

- ready for modeling
- eligible for release
- governable across time

Because Mentim datasets intentionally preserve ambiguity, readiness is not determined by convergence alone.
A dataset is considered ready only when it demonstrates:

- structural integrity
- interpretive traceability
- lifecycle accountability

4.8.1 Dataset Readiness Criteria

A dataset may proceed to modeling, analysis, or publication only when all of the following conditions are satisfied.

1. Structural Completeness

All required fields defined in Section 3 must be present or explicitly set to `NULL` (JSON literal `null`). Keys may not be omitted; missing data must be deliberate and auditable.

- Core Triad ordering must be preserved (State → Event → State).
- Structural & Temporal, Liminality, Energetic & Ethical, Metacognitive, and Administrative layers must be populated wherever applicable.

- All enum values must conform exactly to schema-defined vocabularies.

Constraint: Failure in any structural field disqualifies the batch from release.

2. Distributional Stability

The dataset must report document-level classification distributions for `CLEAR_POSITIVE`, `CLEAR_NEGATIVE`, `AMBIGUOUS`, and `INVERSE_TRANSITION`.

- **Distribution constraints**: No single class may exceed 70% of the dataset unless justified; deviations must be documented in `_dataset_manifest`.
- **Complexity constraint**: High-complexity items ($\mathcal{D} > 0.55$ and $P > 0.60$) must constitute a non-trivial minority (Target: 10–30%). Datasets with 0% high-complexity items are presumed structurally trivial.

3. Reproducible Disagreement

A subset of items must undergo re-annotation. The dataset qualifies only if:

- Disagreement Index ($\mathcal{D}$) and Polarization (P) remain within ±10–15% relative to prior round.
- Dominant `AmbiguityType` assignments persist.
- Confidence scores correlate inversely with measured disagreement.

Note: Variance that collapses or migrates triggers adjudication review.

4. Annotator Cohort Integrity

The dataset must include distributions of `AnnotatorProfile` and `InterpretiveAngle`, rater overlap percentage, and calibration statistics.

- **Cohort constraints**: No single `AnnotatorProfile` should exceed 60% representation.
- Interpretive diversity is treated as a structural requirement, not an optional feature.

Release Gate Condition

Only datasets meeting all four criteria may proceed to model training, public release, research dissemination, or governance reporting. Datasets failing any criterion must be remediated or classified as “Research-Internal Only”.

Governance Principle

Readiness in Mentim does not mean agreement. It means:

- **Structure is intact**
- **Disagreement is patterned**
- **Ambiguity is documented**
- **Variance is reproducible**
- **Cohort composition is transparent**

Release without these safeguards constitutes ontological negligence.

4.8.2 Release Tiers

Mentim defines three release tiers reflecting distinct governance, privacy, and epistemic exposure profiles.

Each tier specifies:

- What data elements are disclosed
- To whom access is granted
- Under what accountability and audit conditions

Release tier designation must be recorded in `_release_manifest` and version-linked to `UAS_Version`.

Tier 1 — Internal Research Release

Scope

Includes full annotation records:

- Core SES fields
- Structural & Temporal layer
- Liminality layer
- Energetic & Ethical fields
- `AnnotatorProfile`
- `InterpretiveAngle`
- `Confidence`
- `AmbiguityReason` (free-text)
- `AmbiguityType`
- `Annotator_ID` (pseudonymous hash)
- Full disagreement metrics ($\mathcal{D}$, P)
- Rater-level variance distributions

Use

- Internal modeling
- Ontology refinement
- IRR diagnostics
- Adjudication review
- Methodological experimentation

Distribution

- Not distributed externally
- Access restricted to authorized internal researchers
- Access logs retained

Governance Rationale

This tier preserves maximal epistemic resolution and interpretive traceability.

It exists to:

- Diagnose schema limits

- Preserve interpretive nuance
- Support ontology evolution

Tier 2 — Controlled External Release

Scope

Intended for:

- Academic collaborators
- Audited institutional partners
- Governance review bodies

Includes:

- Full structural annotations (SES + layered fields)
- `AmbiguityType`
- Aggregated disagreement metrics ($\mathcal{D}$, P)
- Distribution summaries (not rater-level raw variance)

Excludes

- `Annotator_ID` values
- Raw rater-level variance logs
- Free-text `AmbiguityReason` fields

Free-text fields may be released only if:

- Explicitly approved
- Approval logged in `_release_log`
- Reviewer hash recorded
- Timestamp recorded

Governance

- Access governed by data-use agreements (DUA)
- Audit provisions required
- Redistribution prohibited without written authorization

Governance Rationale

Tier 2 balances:

- Interpretive richness
- Research transparency
- Privacy protection

It allows structured disagreement to be studied without exposing individual interpretive traces.

Tier 3 — Public / Open Release

Scope

Includes:

- Schema and codebook

- Version metadata
- Aggregated statistics only
- Class distributions
- Aggregated $\mathcal{D}$ and P summaries

Excludes

- All individual-level annotator metadata
- `AnnotatorProfile` distributions (unless aggregated to cohort-level with $n \geq 5$)
- `InterpretiveAngle` at individual level
- All free-text fields
- Raw item-level annotation records

Disagreement Handling

Disagreement is preserved only at the distributional level.

Privacy safeguard:

- Minimum $n = 5$ per aggregate cell
- Cells below threshold are suppressed or merged

Use

- Benchmarking
- Replication
- Policy citation
- Public transparency
- Governance reporting

Governance Rationale

This tier maximizes:

- Transparency
- Reproducibility
- Public accountability

While enforcing strict de-identification guarantees.

Tier Escalation & Downgrade Rule

Movement between tiers requires:

- Logged approval in `_release_log`
- `UAS_Version` reference
- Named governance reviewer (hashed ID)
- Timestamp

Retroactive expansion of access is prohibited without re-review.

Release Principle

Mentim does not treat release as a binary.

It treats release as a calibrated disclosure decision balancing:

- Interpretive integrity
- Privacy
- Governance risk
- Public accountability

Each tier reflects a different equilibrium point between those forces.

4.8.3 Versioning and Change Control

All released datasets must include explicit provenance metadata sufficient to ensure traceability, reproducibility, and longitudinal comparability.

Required Provenance Fields
Each dataset release must contain:

- `UAS_Version`
- `ProtoTransition` identifier (e.g., PT-01)
- Structured changelog

Changelog Requirements
The changelog must document:

- Definition updates
- Added fields
- Deprecated fields
- Enum modifications
- Anchor revisions
- Backward compatibility notes

The changelog must follow the structured format:
`version | date | change_type | description | breaking_flag`

Where:

- `version` = semantic version identifier
- `date` = ISO 8601 format
- `change_type` = `DEFINITION_UPDATE | FIELD_ADDED | FIELD_DEPRECATED | ENUM_REVISION | STRUCTURAL_CHANGE`
- `description` = human-readable explanation
- `breaking_flag` = `TRUE` | `FALSE`

This format ensures: Human readability, machine parseability, and longitudinal audit compatibility.

Schema Update Policy

When schema updates occur:

- Previously released datasets are not retroactively altered.

- Superseding datasets are issued under new `UAS_Version` identifiers.
- Historical datasets remain valid under their original version.
- **Silent schema drift is prohibited**.

This preserves epistemic lineage and prevents untraceable ontology mutation.

Cross-Version Compatibility

To support longitudinal modeling and cross-version analysis:

- Cross-version mapping tables may be provided.
- If breaking changes are introduced, mapping tables must be published within 30 days of the new version release.
- Mapping tables must explicitly document:
    - Field renaming
    - Enum consolidation or expansion
    - Structural migration rules
    - Deprecated-to-successor relationships

If no valid mapping exists for a breaking change, the incompatibility must be explicitly declared.

This ensures that structural evolution remains accountable and analyzable.

4.8.4 Governance and Access Controls

To ensure responsible use across research, modeling, and public-facing contexts, Mentim datasets are governed by formal access and oversight constraints.

Governance is not advisory; it is structural.

1. Declared Use Constraints

Each dataset release must include a `permitted_use_statement`, specifying whether it may be used for:

- Model training
- Evaluation only
- Interpretability research
- Policy analysis
- Governance benchmarking

Uses outside the declared scope require:

- Explicit approval and logged entry in `_governance_log`
- Requester hash and Reviewer hash
- Documented justification and Timestamp

Constraint: Unauthorized scope expansion constitutes governance violation.

2. Auditability
All datasets must retain traceability to the applicable `UAS_Version` and documentation of:

- Rater protocol (§4.5) and Calibration procedures
- IRR metrics and Disagreement metrics ($\mathcal{D}$, P)
- Aggregated `AmbiguityType` distributions

These materials must be sufficient for independent third-party inspection. **Auditability is mandatory for Tier 2 and Tier 3 releases.**

3. Lifecycle Oversight
Datasets are subject to periodic review to assess:

- Ontological drift (misalignment with updated schema definitions)
- Systematic misuse or scope violations
- Deprecated proto-transition dependencies
- Shifts in interpretive structure over time

Review Cadence: Annual minimum review. Immediate review triggered by `UAS_Version` increment, reported misuse, or regulatory inquiry.

Review Outcomes:

- **Maintained** — continues in active use
- **Superseded** — replaced by a newer version
- **Deprecated** — flagged for limited use
- **Archived** — retained for historical reference

Each action must be accompanied by rationale documentation, version linkage, timestamp, and governance reviewer hash.

Governance Principle

Mentim datasets are treated as evolving epistemic artifacts. Governance ensures that:

- Structural integrity persists across time
- Ontological change remains visible
- Interpretive variance remains preserved
- Use remains aligned with declared epistemic purpose

Versioning protects lineage. Governance protects meaning.

4.8.5 Governance Statement

Mentim treats datasets as living governance artifacts, not static training corpora.
Annotated data within the Unified Annotation Schema is understood as an evolving epistemic record—one that documents not only what was labeled, but how meaning was interpreted, where disagreement emerged, and under which structural conditions the

interpretation occurred.

By enforcing readiness standards that privilege structural integrity over forced consensus, and release controls that preserve interpretive traceability, UAS ensures that annotated datasets remain:

- **Ethically accountable** — moral drivers and interpretive tensions are explicitly encoded rather than abstracted away
- **Epistemically transparent** — disagreement, ambiguity, and confidence are preserved as measurable signals
- **Structurally reproducible** — versioning, provenance, and governance controls prevent silent ontology drift
- **Fit for modeling uncertainty** — variance is retained as first-class data rather than collapsed into majority labels
- **Suitable for regulatory, research, and audit scrutiny** — lineage, scope, and lifecycle controls are documented and inspectable

This completes the annotation lifecycle:
From human interpretation
→ to structured capture
→ to measured disagreement
→ to adjudicated refinement
→ to governed release

In Mentim, governance is not an external compliance layer. It is embedded within the ontology itself.

5. The Path Forward

This document—Unified Annotation Schema (UAS) Codebook, Version 1.1—formalizes the first operational unit of meaning within the Mentim framework.

Proto-Transition 01 (PT-01): The Shattering of Innocence constitutes the initial mapped transformation in what is intended to become a structured atlas of human awareness. It encodes the moment when unexamined coherence yields to ethical rupture, uncertainty, and the emergence of inquiry.

PT-01 is intentionally narrow in scope. Its function is not to exhaustively model cognition, but to demonstrate that qualitative transitions of awareness can be:

- specified,
- annotated,
- measured,

- and analyzed with rigor,

without collapsing ambiguity or suppressing interpretive variance.

This codebook therefore serves as a proof of method for a broader objective:
the development of a structured science of transformation — a system capable of modeling how perception, affect, ethics, and identity reorganize through experience and time.

Next Steps

1. Conduct the Initial IRR Study
Apply this codebook to generate an initial dataset of ≥ 50 annotated PT-01 examples, using the rater protocols, overlap requirements, and reporting standards defined in Section 4.
The objective is not convergence, but structural validation.

2. Analyze Disagreement as Structure
Compute IRR, Disagreement Index ($\mathcal{D}$), and Polarization (P) at both document and span levels. Interpret divergence as signal:

- Refine definitions where schema gaps are revealed.
- Preserve variance where multivalence is legitimate.
- Document all revisions under formal version control.

Convergence is diagnostic. Variance is epistemic.

3. Public Release and Citation
Publish the annotated dataset (under defined release tier), reporting package (IRR + complexity metrics), and UAS v1.1 codebook. This establishes the Unified Annotation Schema as a citable standard.

4. Design Proto-Transition 02 (PT-02)
Extend the framework to **PT-02: The Seeker's Resolution**. PT-02 will model stabilization, integration, or reframing following disillusionment. This transition tests:

- continuity across rupture,
- coherence restoration,
- and longitudinal identity restructuring.

Together, PT-01 and PT-02 initiate the first two segments of a multi-transition cognitive arc.

Developmental Adequacy and Future Schema Extensions

As the framework evolves from isolated transitions toward longitudinal arcs, future versions of UAS may incorporate formal mechanisms for assessing developmental adequacy across sequences of Narr-Atoms.

Building on the conceptual foundation of **Developmental State → Event → State (D-SES)**, such extensions would enable evaluation of:

- expectation formation,
- salience differentiation,
- disciplined confidence modulation,
- cross-transition coherence stability.

Importantly:

- These mechanisms are intended as optional certification-level overlays, preserving backward compatibility of the core SES grammar.
- UAS v1.0 remains fully sufficient for modeling rupture, uncertainty, and the onset of inquiry.
- Developmental adequacy becomes relevant only when transitions are evaluated in aggregate — across time, contradiction, and recovery.

Future work may therefore include:

- defining optional overlay fields for developmental signals,
- establishing longitudinal calibration procedures,
- articulating criteria for `Narr-Atom-D` designation,
- formalizing sequence-level structural coherence metrics.

The objective is not to accelerate systems toward premature certainty. It is to define the structural conditions under which confidence, responsibility, and coherence are earned.

Closing Statement

The Unified Annotation Schema does not attempt to resolve meaning into certainty. It renders transformation legible.

By treating:

- disagreement as structured data,
- uncertainty as measurable signal,
- and transitions as first-class analytic objects,

UAS establishes the foundation for accountable, introspective systems — human or artificial — capable of change without incoherence.

This is the first operational map. The atlas expands from here.

Appendix A — How to Cite UAS Datasets

This appendix defines the canonical citation standard for datasets produced using the Mentim Unified Annotation Schema (UAS). Its purpose is to ensure reproducibility, traceability, and governance-grade attribution across academic, regulatory, and commercial contexts.

A.1 Why Citation Matters in UAS

UAS datasets are not static label corpora. They are:

- Interpretive artifacts capturing human meaning under uncertainty
- Versioned governance objects tied to a specific ontology state
- Distributional records where disagreement is signal, not error

As such, proper citation must preserve:

- Schema version
- Proto-Transition scope
- Annotation philosophy (non-convergent)
- Release tier and governance constraints

Citation in UAS is therefore not clerical—it is epistemic attribution.

A.2 Required Citation Elements

Every citation of a UAS dataset must include:

- Dataset Title
- Unified Annotation Schema Version
- Proto-Transition(s) Included
- Release Tier
- Publisher
- Year
- Persistent Identifier (DOI, registry ID, or hash)
- Governance Note (non-convergent annotation)

Note: Citation lacking any required element does not qualify as governance-grade and may not preserve reproducibility or audit eligibility.

A.3 Canonical Citation Format (Academic)

APA-style (recommended)

Mentim AI. (2026). Unified Annotation Schema Dataset: Proto-Transition 01 – The Shattering of Innocence (UAS v1.0) [Annotated dataset, Tier 3 Public Release]. https://doi.org/XXXX

Replace `XXXX` with the registered DOI upon release; preprints must use the registry identifier format `Mentim.YYYY.NNNNN`.

BibTeX

```
@dataset{Mentim_uas_pt01_2026,
  author       = {Mentim AI},
  title        = {Unified Annotation Schema Dataset: Proto-Transition 01 -
                  The Shattering of Innocence},
  year         = {2026},
  version      = {UAS v1.0},
  publisher    = {Mentim AI},
  note         = {Non-convergent, complexity-aware annotation schema},
  doi          = {XXXX}
}
```

BibTeX

```
@dataset{Mentim_uas_pt01_2026,
  author       = {Mentim AI},
  title        = {Unified Annotation Schema Dataset: Proto-Transition 01 -
                  The Shattering of Innocence},
  year         = {2026},
  version      = {UAS v1.0},
  publisher    = {Mentim AI},
  note         = {Non-convergent, complexity-aware annotation schema},
  doi          = {XXXX}
}
```

A.4 Canonical Citation Format (Regulatory / Audit Use)

For compliance filings, risk assessments, or governance reports:

> *"Analysis incorporated data annotated under the Mentim Unified Annotation Schema (UAS v1.0), Proto-Transition 01 (The Shattering of Innocence), using a non-convergent, complexity-aware methodology. Dataset release tier: Controlled External. Schema and annotation lineage available for audit."*

This phrasing explicitly signals:

- Interpretive intent
- Version lock
- Governance suitability

A.5 Canonical Citation Format (Model Cards & Technical Reports)

For inclusion in model cards or evaluation sections:

> *"Training and evaluation incorporated human-annotated narrative transitions labeled using the Mentim Unified Annotation Schema (UAS v1.0). Annotations preserve interpretive distributions rather than enforcing consensus; disagreement metrics ($\mathcal{D}$, P) were retained and modeled."*

A.6 Citing Subsets, Filters, or Derived Corpora

When using a subset or derived dataset, include:

- Original dataset citation
- Selection criteria
- Whether disagreement metrics were preserved or collapsed

Example:

"Derived from Mentim AI (2026), UAS v1.0 PT-01 dataset. Subset restricted to `CLEAR_POSITIVE` *and* `AMBIGUOUS` *items with* $\mathcal{D} > 0.55$*. Original interpretive distributions preserved."*

Constraint: Minimum $n = 50$ per derived corpus; document justification if smaller.

A.7 Prohibited Citation Practices

The following practices are not permitted when citing UAS datasets:

- Claiming annotations represent a single "ground truth"
- Omitting schema version or proto-transition
- Treating disagreement as labeling error
- Releasing annotator-level metadata outside the declared release tier

Violations invalidate governance guarantees and audit eligibility.

A.8 Governance Statement (Citation-Level)

All UAS datasets are governed artifacts. Citation implies acceptance of the schema's interpretive philosophy: **ambiguity, disagreement, and liminality are first-class signals of meaning**.

Appendix Summary

This citation standard ensures that UAS datasets remain:

- Citable without distortion
- Auditable without IP leakage
- Comparable across time
- Legible to researchers, regulators, and systems alike

It completes the transition from annotation as labeling to annotation as epistemic infrastructure.