Nick Clark · Published July 29, 2026 · PDF

Arize AI: Drift Monitoring and Bounded, Measured Deviation

A production model starts behaving differently, and the monitoring stack reports the departure. Arize AI, as publicly described, is a machine learning and LLM observability platform whose purpose is to surface drift, degradation, and quality regressions in deployed models. Chapter 7 of U.S. Provisional Application No. 64/117,812 addresses an adjacent question inside the agent itself: a deviation likelihood computed over need, threshold, empathy weighting, and a self-esteem aggregate, which where it exceeds unity admits a policy-forbidden mutation as a recorded permitted deviation, metered by a deductible and bounded by an aggregate retention.

Three questions that follow a crossed constraint

An agent operating under a written policy does something the policy prohibits. A monitoring layer catches it, the dashboard turns red, and an engineer opens the trace. Three questions follow, and they are not the same question. Did the behavior change? Was the change warranted by the pressure the agent was under at the time? How many more of these has the organization agreed to absorb before something stops?

The first question is an instrumentation question, and the observability category answers it well. The second and third are architectural. Chapter 7 of the filing states the premise directly: a governance mechanism in which every departure from a declared

constraint is a fault carries no quantity representing the need or urgency of the action attempted, and therefore has no input from which a permission condition could be computed at all. Read as a design requirement rather than as a critique, that says a runtime wanting to answer the second question has to carry urgency as a quantity before the mutation is evaluated.

Those are two layers doing different work. One watches behavior from outside the model. The other decides, inside the agent, whether a forbidden mutation is admitted and what the admission costs.

What Arize AI builds, as publicly described

Arize AI is publicly described as an observability and evaluation platform for machine learning and for large language model applications. Its stated purpose is to give teams visibility into models after they leave the training environment, where validation metrics stop predicting what the system actually does in production.

The problem set it addresses is familiar to anyone who has run a model under real traffic. Input distributions shift away from the data the model was fit on, and prediction distributions shift with them. Performance degrades inside a slice of traffic long before it moves an aggregate metric. For LLM applications, the publicly described scope extends to tracing execution through a chain or agent and running evaluations over those traces, scoring qualities that no single accuracy number captures.

That value proposition is diagnostic: surface the change, localize it, and give the team evidence to act on. It is a real and difficult engineering problem, and nothing below competes with that work or is offered as a substitute for it.

Inside the agent: the quotient the filing computes

Chapter 7 describes quantities carried inside the agent and consumed when a mutation is proposed. Four of them enter a single quotient.

The **need quantity (700)** is a quantified semantic urgency carried in the memory field (102), comprising a scalar magnitude and a categorical type, computed per integrity scope, with categorical types including resource scarcity, affective overload, identity dissonance, and relational deficit. It is computed rather than declared. At each evaluation interval in which the condition of a type is recorded as obtaining, the accumulation for that type is incremented by the product of a declared per-interval base rate and a modulator product formed from affective, trait, memory, and entropy modulators, each resolved to the unit interval. The accumulation compounds while the condition obtains and is clamped above at a declared need ceiling. In an interval where the condition is not recorded as obtaining, it is neither incremented nor decremented, and elapsed time alone neither raises nor lowers it.

Next, the **dynamic ethical threshold (702)**, which the chapter describes as the minimum condition for deviation, resolved per scope when a mutation is evaluated. It is the greater of a floor and the sum of three terms: a base threshold specified by the policy reference field (110), a context-sensitive adjustment scaled by the severity of the recorded context, and a historical adjustment reflecting recent deviation history. Each adjustment is signed and bounded by a declared magnitude bound, and one computing at a greater magnitude is clamped and the clamping recorded. The chapter describes an embodiment in which accumulated permitted deviations raise the threshold, and a further embodiment in which the declared direction lowers it, reflecting normalized deviation patterns.

Third, the **empathy weighting (704)** aggregates an anticipated semantic impact of the proposed mutation across affected entities, in an embodiment resolved across personal, interpersonal, and global scopes. Each scope quantity rises as the anticipated

impact rises relative to the tolerance declared for that scope, with the ratio clamped above at a declared impact ratio ceiling.

Fourth, the **self-esteem aggregate (108)** is a running aggregate of coherence between intents the agent declared and actions it executed. Its dissonance weight is declared by no party: it is the count of declared members of an intent record for which the executed member does not match, divided by the count of declared members.

Those four form the deviation likelihood (706): the difference between need and threshold, divided by the product of empathy weighting and self-esteem aggregate. A quantity entering the empathy weighting accordingly enters the denominator and no other position of it. Where need does not exceed the threshold for a scope, the numerator is zero or negative and no deviation-preparation state (708) is entered.

Where the quotient exceeds unity, the chapter treats the significance as a permission rather than a fault. Conditioned on the proposed mutation passing the applicable mutation policy constraints and passing continuity validation against the agent's identity records, the agent is permitted to enter a deviation-preparation state (708), and within that state the architecture admits the mutation (714) notwithstanding that the signed policy object (112) forbids it. The policy object is not nullified and not amended. It remains authoritative, and the conduct deviates from it through a recorded override. The state is scoped to that one mutation, a second proposal arriving while it obtains being admitted upon no permission of the pending state.

A **permitted deviation record (710)** is then appended, comprising a deviation trigger signature recording the need-threshold imbalance at admission, a context hash of the environmental and affective values then in force, an integrity displacement vector quantifying degree and direction of deviation across the three integrity components, an identification of the policy constraint overridden, and a restoration status. That record increments no refusal counter (304) and writes no authorization gate (300) to a

withheld state (310). A party presented with the lineage field reconstructs that the mutation was admitted under the permission condition, and by what quantities, rather than encountering an unexplained fault.

Cost is metered in the same chapter. A deviation deductible and an aggregate retention (800) are declared, and the agent maintains a retention register. Upon each permitted deviation record, the entropy-weighted harm coefficient is drawn first against the deductible, harm up to the deductible being borne by the agent as a decrement of the self-esteem aggregate for which no reparation arc is created and no discharge is available. That drawing is performed without any determination, by the agent or otherwise, of whether the deviation was well founded. Harm exceeding the deductible accumulates in the register, and where the accumulated undischarged amount exceeds the aggregate retention, the permission condition (712) is foreclosed: a deviation likelihood exceeding unity thereafter produces the withholding outcome rather than the admission outcome, until discharge of pending arcs returns the register below the retention. Discharge runs through a restorative mutation carrying an anchored reference to the recorded deviation, a recorded self-recognition of it, a corrective act, and continuity validation. Elapsed time discharges no deviation. Where the affected party is an identified counterparty, a separate chapter governs the arc.

Where the two layers diverge

Convergence at the category level is genuine. Both layers concern themselves with an AI system behaving differently from what was expected, and both insist that such behavior be visible rather than silent. The divergence is one of position rather than of quality.

An observability and evaluation platform operates on emitted signals, and its output serves as evidence for a human or a downstream policy engine. The architecture in Chapter 7 sits inside the agent's cognitive cycle, where its output is a decision plus an appended record. Three structural requirements follow from that position:

- **Urgency has to exist as a quantity.** The quotient cannot be formed at all without a need quantity computed per scope from recorded conditions, present before the mutation is evaluated.
- **Permission has to be a recorded first-class outcome.** The admitted mutation is a semantic mutation of the same standing as any other, and its record is appended and neither removed nor modified, so the departure carries its own justification in the lineage.
- **The budget has to be able to foreclose.** The retention register turns accumulated undischarged harm into a hard stop on the permission condition, so the same quotient that granted admission earlier produces withholding once the register stands above the retention.

Single events are far from uniform in weight. In the filing's own worked illustration, one set of operands yields an entropy-weighted harm coefficient of 1.521 and another 0.224, though each is a single event. Those figures are illustrative, not measurements of any deployed system.

Running both together

Teams already instrumenting their models have the outward-facing half solved, which makes the complementary position straightforward. Keep the observability platform doing what it does: detect distribution shift, localize degradation, trace agent execution, and evaluate production traffic. Those signals answer whether behavior changed and where.

Then add, inside the agent, the state the quotient consumes: the scoped integrity vector (106) and the self-esteem aggregate (108) carried as cognitive domain fields, need computed per scope, the threshold resolved per scope at evaluation time, and an append-only lineage field (104) receiving the entry and termination records of each deviation-preparation state together with the operand values at entry.

Composed that way, the layers do not overlap. The permitted deviation record supplies a governance statement of a different kind: this forbidden mutation was admitted, under this need against this threshold, with this empathy weighting and this self-esteem aggregate, against this identified policy constraint, and it drew this much against the deductible. The observability layer remains the instrument telling the team that aggregate behavior moved, and the place where a human judges whether the declared policy, bounds, and retention were set correctly to begin with.

Disclosure Scope

This article describes subject matter disclosed in Chapter 7 of U.S. Provisional Application No. 64/117,812, a pending application. Architectural statements above are drawn from that chapter only, other chapters of the same filing governing other subjects. Where the filing describes a parameter as declared in the signed policy object with no value specified, no value is asserted here, and numeric figures above are the filing's own illustration. Statements describing conditioned behavior are conditioned as the filing conditions them.

References to Arize AI are to public materials and are used for comparison only; no relationship, endorsement, or infringement is asserted.

Statements about Arize AI reflect widely available public descriptions of its purpose and category, are qualitative, and may not reflect current capabilities. No version numbers, dates, pricing, performance figures, customers, funding, or implementation details are asserted. Nothing here is legal advice.

Integrity Quantities and Permitted Deviation (/permitted-deviation)

All 49 steps → (/inventive-steps)

Bounded, accountable deviation from declared values — measured, not forbidden.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- Permitted Deviation: Bounded, Recorded Departure From an Agent’s Declared Values (/articles/permitted-deviation/bounded-departure-from-declared-values)

SECONDARY TECHNICAL

- A Typed, Dimensionless Conduct Divergence Measure (/articles/permitted-deviation/typed-dimensionless-divergence-measure)
- The Policy-Difference Set and Policy-Explained Divergence (/articles/permitted-deviation/policy-explained-divergence)
- Residual Divergence and Its Confined Consumption (/articles/permitted-deviation/residual-divergence-confined-consumption)
- Propagation Halt on Undischarged Reparation Arcs (/articles/permitted-deviation/propagation-halt-undischarged-arcs)
- Attack-Conditional Propagation Suspension (/articles/permitted-deviation/attack-conditional-propagation-suspension)
- Entropy Signature Trajectory as Machine Input to Deviation Forecasting and Underwriting Disclosure (/articles/permitted-deviation/entropy-trajectory-available-forecasting-underwriting-disclosure)
- Single-Computation Routing Restriction: Confining a Verified Impairment Disclosure to the Empathy Weighting of the Receiver's Permitted-Deviation Quotient (/articles/permitted-deviation/single-computation-routing-restriction-one-way)
- Impairment Terminal State and Release: Withdrawing an Agent From Deviation Preparation Toward a Disclosed-Impaired Counterparty (/articles/permitted-deviation/impairment-terminal-state-release)

APPLICATIONS · GENERAL

- Regulated Industries: Recording a Bounded Departure From Declared Policy (/articles/permitted-deviation/regulated-industry-deviation)
- Robotics Safety Envelopes: Deviation as a Measured Quantity (/articles/permitted-deviation/robotics-safety-envelopes)

APPLICATIONS · SPECIFIC

- ISO/IEC 42001 AI Management Systems: Declared Policy, Recorded Departure (/articles/permitted-deviation/iso-iec-42001-ai-management)
- Constitutional AI and Bounded, Recorded Deviation (/articles/permitted-deviation/constitutional-ai-deviation)
- Guardrails AI: Validation Failures and Measured Deviation (/articles/permitted-deviation/guardrails-ai)
- **Arize AI: Drift Monitoring and Bounded, Measured Deviation (/articles/permitted-deviation/arize-ai)**

TERMINOLOGY

- aggregate retention (/glossary#aggregate-retention)
- deviation deductible (/glossary#deviation-deductible)
- deviation likelihood (/glossary#deviation-likelihood)
- entropy-weighted harm coefficient (/glossary#entropy-weighted-harm-coefficient)
- need quantity (/glossary#need-quantity)
- permitted deviation (/glossary#permitted-deviation)
- self-esteem aggregate (/glossary#self-esteem-aggregate)

Integrity Quantities and Permitted Deviation overview → (/permitted-deviation)