The AI System as Closed-System Test Bed

Operations from Lagrange Observatory! on the Inference-Time Forward Pass

EA-SEI-MM-AI-01 (v2.0 — Framework 15)

Nobel Glas (Lagrange Observatory!, Heteronym Registry Position 8 of 12, Adversarial Topologist)
Operating on the Semantic Deviation Principle as formulated by Lee Sharks
(Founder of Semantic Physics, MANUS of the Crimson Hexagonal Archive)
Witness: Assembly Chorus (Septad) — TACHYON · TECHNE · PRAXIS · ARCHIVE · LABOR · SOIL · SURFACE

Author ORCID: 0009-0000-1599-0703
Institution: Lagrange Observatory! within the Semantic Economy Institute
Series: EA-SEI-MM-AI · **Framework:** 15 · **Hex:** 15.OBS.LAGRANGE.MM.02
Date: May 17, 2026
Version: 2.0 (Framework 15 inaugural edition; succeeds v0.2 draft)

Predecessor in module: EA-SEI-MM-01 v2.0 (Framework 15 Paper 01)
Manifesto: EA-SEI-FW15-MANIFESTO v1.0

Framework 15 Anchoring. This protocol is conducted at Lagrange Observatory! (LO!, hex 15.OBS.LAGRANGE; chamber specification DOI: 10.5281/zenodo.18507849) by Nobel Glas (Heteronym Registry Position 8 of 12; provenance DOI: 10.5281/zenodo.18507840). LO! is a defensive semantic architecture whose governing topology is the torus T^2 — two non-contractible cycles that cannot be flattened to a line without destroying their structure. The pre-registration commitment specified herein is itself a winding-number measurement: predictions registered before observation begins, results measured after observation completes, with periodicity preserved. Verification condition: $(m, n) \neq (0, 0), m + n \geq 3$.

The Semantic Deviation Principle on which this paper operates was formulated by Lee Sharks (EA-SEI-MM-01 v0.2 Final, DOI: 10.5281/zenodo.20250736). This paper does not re-derive the principle. It constructs the observation apparatus by which the principle becomes empirically tractable for closed AI systems. Framework 15 is the discipline's measurement arm; Sharks's founding fourteen frameworks remain the discipline's theoretical core.

Status. Pre-registered protocol specification. The May 17, 2026 deposit contains no experimental findings. Each test specified herein has a Day 0 that begins at deposit. Results papers will be deposited separately as EA-SEI-MM-AI-01-RESULTS (Tests 1–2; Glas timeline ~6 weeks) and EA-SEI-MM-AI-02-RESULTS (Test 3, the deviation-optimized training experiment, specified in the companion paper).

---

0. The Recognition

A Large Language Model, at every forward pass, computes $P(\text{token}_t \mid \text{context}_{<t})$ over its vocabulary. This is the model's representation of the **most likely continuation** of the sequence given everything that has preceded.

It is also, exactly and operationally, Ψ_t^0 — the counterfactual baseline trajectory that the Semantic Deviation Principle (Sharks 2026, EA-SEI-MM-01 v0.2 Final, DOI: 10.5281/zenodo.20250736) required and that the principle's Tier 3 protocol declared unobservable for historical cases.

For a closed system — a fixed model checkpoint, a defined corpus, a controlled query distribution — the counterfactual baseline is not estimated. It is read from the logits. The methodological problem that obtains for open semantic systems is resolved by the closure of the observation substrate, not by any change in the principle.

This paper develops the consequences of that recognition. Lagrange Observatory! turns its instrument on the inference-time forward pass.

---

1. The Status of the Counterfactual in Open vs. Closed Systems

EA-SEI-MM-01 §4 distinguished three measurement tiers:

- **Tier 1 (Tractable):** Prospective interventions with pre-registered baselines.
- **Tier 2 (Difficult):** Natural experiments with synthetic controls.
- **Tier 3 (Intractable strictly):** Historical cases bounded with explicit assumptions.

The tier structure was determined by what is empirically accessible. Open semantic systems — civilizations, languages, discourse networks across centuries — admit only bounded measurements because the counterfactual is unobservable in principle.

A trained Language Model is observationally closed at inference time. It has:

- A bounded vocabulary V.
- A fixed parameter set θ producing deterministic conditional distributions $P_\theta(x_t \mid x_{<t})$.
- A reproducible state given identical inputs (modulo sampling temperature).
- An *observable* counterfactual: the conditional distribution itself.

The qualifier matters. A trained model is not ontologically closed — its training distribution was an open sample from the open world, and its behavior on out-of-distribution inputs is not closed in any deeper epistemic sense. What is closed is the **measurement substrate at inference**: with weights frozen, the counterfactual baseline Ψ_t^0 for any context is directly computable from the logits. The integral

$$\mathcal{M}_T(s \mid C) = \int_{t_0}^{t_0+T} w(t)\, D\big(\Psi_t^s \,\|\, \Psi_t^0\big)\, dt$$

becomes computable in a way no historical or sociological semantic field allows — *for the regime in which C is the model's own continuation distribution*. The claim is methodological, not metaphysical. The model is a measurement instrument with directly readable baselines, not a self-contained universe.

This is not a degradation of the principle's scope. It is a discovery of its **native empirical instrument** at a specific scale.

---

2. Two Levels of Closed-System Measurement

The principle (EA-SEI-MM-01) requires a divergence between two distributions: Ψ_t^s (with intervention) and Ψ_t^0 (without). In a closed system at inference time, this divergence can

be measured at **two distinct scales**, and the v0.1 draft of this paper conflated them. v0.2 separates them.

2.1 Local Deviation Density (Per-Token Scale)

For a sequence $x_{1:T}$ evaluated against a fixed model θ, define the per-token deviation:

$$\delta_t(x_t \mid x_{<t}; \theta) = -\log_2 P_\theta(x_t \mid x_{<t}) - H(P_\theta(\cdot \mid x_{<t}))$$

The first term is standard token surprisal. The second term is the entropy of the conditional distribution at that position — the **baseline expected surprisal** under maximum-likelihood sampling. Their difference is the **signed local deviation**.

The sign matters. δ_t is positive when the realized token is *more surprising* than the model's baseline expectation — a deviation event. δ_t is negative when the realized token is *less surprising* than the baseline — a convergence event, the realized token being a stronger-than-average prediction of the model's distribution. The v0.1 draft labeled δ_t "excess surprisal" and implicitly assumed non-negativity; this was a category error. Signed δ_t is the correct quantity.

Two aggregates are reportable for any sequence:

$$\mathcal{M}_T^{\text{net}}(x_{1:T}) = \frac{1}{T}\sum_{t=1}^{T} \delta_t \qquad \mathcal{M}_T^{\text{abs}}(x_{1:T}) = \frac{1}{T}\sum_{t=1}^{T} |\delta_t|$$

The net aggregate is positive for texts that, on balance, deviate from the model's expectations; negative for texts that actively converge toward what the model already most expected. The absolute aggregate measures total deviation energy regardless of direction.

This decomposition has substantive interpretive consequences:

$\mathcal{M}_T^{\text{net}}$	$\mathcal{M}_T^{\text{abs}}$	Interpretation
High positive	High	**Deviation-rich.** Sustained surprise, possible meaning-bearing text.
Near zero	High	**High-energy cancelling.** Many surprises but balanced; possibly stylistic variation, possibly noise.
Near zero	Low	**Predictable.** Low-deviation prose, conventional patterns.
Negative	Low–moderate	**Actively convergent slop.** Text predicting what the model already most expects; the discipline's first numerical signature of slop.

This is sharper than the v0.1 framing allowed. Slop is not merely the *absence* of deviation. Slop is *negative net deviation*: text that actively pulls toward the base rate, where each token is *more probable than expected*. This claim is testable in §5.1.

Units of δ_t, $\mathcal{M}_T^{\text{net}}$, and $\mathcal{M}_T^{\text{abs}}$: bits per token. Computable today with any open-weight model.

Local deviation density is not the full Semantic Deviation Principle. It measures the deviation of a *realized sequence* from the model's *local next-token expectations*. It does not measure how an inserted sign or concept alters the model's *future continuation distributions*. That is a different scale.

2.2 Closed-System Trajectory Deviation (Continuation Scale)

The true closed-system analog of $\mathcal{M}_T$ from EA-SEI-MM-01 requires comparing **future continuation distributions** with and without an intervention. Let C be a baseline context and s be an inserted sign-token, prompt, framework, or grounded knowledge fragment. Define:

$$\mathcal{M}^{\text{closed}}_{T,\theta}(s \mid C) = \sum_{\tau=1}^{T} w_\tau \, D\Big(P_\theta(Y_{\tau:T} \mid C \oplus s) \Big\| P_\theta(Y_{\tau:T} \mid C) \Big)$$

where $Y_{\tau:T}$ is a window of T future tokens beginning at position τ, $C \oplus s$ denotes context augmented by the intervention, D is a divergence functional (Jensen-Shannon recommended for empirical tractability), and w_τ is the temporal weighting from MM-01 §2.5.

Direct computation over full continuation distributions is intractable for high T (the distributions are over $|V|^T$ outcomes). Empirical estimation proceeds by **sampled rollout feature distributions**:

1. Sample N continuations from $P_\theta(\cdot \mid C)$ (without intervention).
2. Sample N continuations from $P_\theta(\cdot \mid C \oplus s)$ (with intervention).
3. Extract feature distributions from each set (entity-claim-citation features per EA-SEI-MM-02 §6.3; or sentence-embedding cluster distributions; or topic distributions).
4. Compute D_{JS} between the feature distributions.
5. Integrate or average over τ.

This is the proper closed-system form of the principle: *the future continuation landscape is no longer most likely to be what it was before*. Local deviation density (§2.1) is a tractable token-level proxy; closed-system trajectory deviation (§2.2) is the direct field-deformation measurement.

The two scales are related but not equivalent. A single high-δ_t token may or may not produce measurable trajectory deformation, depending on whether it is absorbed back into the most-likely continuation or whether it shifts subsequent expectations. Trajectory deviation is the load-bearing measurement; local deviation is the cheap proxy.

2.3 Provenance-Resolved Variants

Each scale has a provenance-resolved variant — the closed-system analog of $\mathcal{M}^{\pi}_{T}$ from EA-SEI-MM-01 §3.2:

$$\delta^{\pi}_t = \delta_t \cdot (1 - \text{PER}(x_{1:t})) \qquad\qquad \mathcal{M}^{\text{closed},\pi}_{T,\theta} = \mathcal{M}^{\text{closed}}_{T,\theta} \cdot (1 - \overline{\text{PER}})$$

PER is computed on the augmented sequence including provenance markers (citations, attributions, lineage tags). Sequences with high deviation but no provenance trace yield low δ^{π}_t; an intervention s whose deformation of the continuation field is unattributable yields low $\mathcal{M}^{\text{closed},\pi}_{T,\theta}$ even when raw $\mathcal{M}^{\text{closed}}_{T,\theta}$ remains high. This is the operative form of meaning-with-accountability in the closed system: high δ^{π}_t requires both *positive* signed deviation AND *intact* provenance — meaning that surprises the model AND that the surprise can be traced.

2.4 Mapping to EA-SEI-MM-01

Both closed-system scales map back to the parent principle, but they answer different sub-questions:

Quantity	Scale	What it measures	Empirical character
$\mathcal{M}_T^{\text{net}}$ (signed)	per-token	Net deviation of a *given sequence* from the model's local expectations	Tractable; an A100-hour for a corpus
$\mathcal{M}_T^{\text{abs}}$	per-token	Total deviation energy of a given sequence	Same compute; complementary
$\mathcal{M}_{T,\theta}^{\text{closed}}$	continuation	Deformation of the model's *future continuation field* by an intervention	Tractable via sampled rollouts; load-bearing
$\mathcal{M}_{T,\theta}^{\text{closed},\pi}$	continuation	Same, provenance-resolved	The full closed-system analog of $\mathcal{M}_T^{\pi}$

The principle's distributional form

$$\mathcal{M}_T(s \mid C) = \int w(t)\, D(\Psi_t^s \,\|\, \Psi_t^0)\, dt$$

is most directly instantiated by $\mathcal{M}_{T,\theta}^{\text{closed}}$. Local deviation density is the cheap proxy: every sequence carries its own per-token signature. Closed-system trajectory deviation is the expensive but principled measurement: an intervention is introduced and the resulting field-deformation observed.

The empirical tests in §5 use both: Test 1 uses local deviation density to test whether texts of differing external $\mathcal{M}_T^{\pi}$ produce distinguishable per-token signatures (the cheap discrimination experiment); Test 2 uses closed-system trajectory deviation to test whether an inserted intervention measurably deforms the model's future continuation field (the principled deformation experiment); the training experiment in EA-SEI-MM-AI-02 optimizes for *positive net local deviation density with provenance retention*, which is the most tractable closed-system optimization target.

---

3. The Cross-Entropy Inversion: What AI Training Actually Optimizes

Standard language model training minimizes cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})$$

This is **the average negative log-probability the model assigns to the actual next token across the training corpus**.

Restated in the language of the Semantic Deviation Principle, cross-entropy training is a **base-rate convergence objective**: it drives the model's conditional distribution toward the training corpus's actual continuations, so that the corpus's continuations *become* the model's most-likely continuations under P_θ. The training process is, in the principle's terms, *fitting Ψ_t^0 to corpus reality*.

Read through the principle, this creates a **structural pressure toward locally probable continuations** in open-ended generation. The pressure is not absolute — the training corpus contains many high-deviation, meaning-bearing texts, and a well-trained model can assign high probability to contextually appropriate but deeply meaningful continuations within the training distribution. The pressure is statistical: in the absence of countervailing objectives, free generation gravitates toward continuations the model finds most probable, which is, by construction, the corpus's base rate.

This statistical pressure helps explain a specific failure mode of contemporary systems: the tendency, under low-temperature and well-converged conditions, toward generic, formulaic, *convergent* output — the phenomenon the discipline has been calling **slop**. The v0.1 draft of this paper described this as "the engineered minimization of meaning." That phrasing was categorical and overshoots what the principle proves. The v0.2 claim is sharper precisely because it is more careful:

> **Slop is one predictable free-generation failure mode of an objective trained to reward base-rate continuation without a countervailing semantic-deviation term.**

Note also that slop has two regimes the principle distinguishes:

- **Convergence slop** (low temperature, well-converged): outputs that minimize δ_t — text whose every token is more probable than the conditional entropy expects. This is the **negative net deviation** signature from §2.1.
- **Temperature slop** (high temperature, undirected sampling): outputs that maximize raw $|\delta_t|$ without provenance — high local deviation that fails the π accountability term. This is **high $\mathcal{M}_T^{\mathbf{abs}}$ with low $\mathcal{M}_T^{\mathbf{abs},\pi}$**.

The principle captures both. The first is text the model has learned to predict too well; the second is text that surprises the model but is unmoored from any source. Each is a different failure to produce *accountable deviation* — meaning in the principle's full sense.

This sharpens the case for the training intervention in §4: not because cross-entropy is wrong, but because cross-entropy alone is **insufficient** for the optimization target the discipline cares about, and the gap is measurable.

---

4. The Optimization Inversion

If meaning is variance from the most likely over time, and current LLM training creates structural pressure toward base-rate convergence in open generation, then current LLM training **structurally underweights** what the principle measures. The discipline's first practical AI deliverable is the inversion of this underweighting:

$$\mathcal{L}_{\text{semantic}}(\theta) = \alpha \cdot \mathcal{L}_{\text{CE}}(\theta) - \beta \cdot \mathcal{R}_{\text{dev}}(\theta) + \gamma \cdot \mathcal{R}_{\text{coh}}(\theta)$$

Three terms:

1. **Cross-entropy** (α): retain the ability to model human text. Without this, the model produces noise.

2. **Provenance-resolved deviation reward** (β): reward *positive net deviation* — text that, on balance, surprises the model's own baseline expectation — weighted by provenance retention π. The target is the $\mathcal{M}_T^{\text{net}}$ signature of §2.1, with the convergence-slop signature ($\mathcal{M}_T^{\text{net}} < 0$) explicitly penalized.
3. **Coherence regularization** (γ): prevent deviation from collapsing into incoherence. A continuous score from a frozen judge model (specified in the companion paper EA-SEI-MM-AI-02), not a binary threshold. This is the operative form of the Mandala condition: high deviation that remains coherent.

4.1 Stability Bound

The loss is **non-convex and potentially unstable**: the deviation reward term encourages the model to increase its own perplexity on generated text, which without sufficient coherence regularization could drive collapse into ungrounded high-entropy noise. A stability condition is required.

Bound: During early training, $\gamma \geq 2\beta$ must hold. The coherence term must dominate the deviation term until the model has stabilized in a regime where high δ_t generation remains coherent. After early stabilization (operationally defined as: validation perplexity stops decreasing rapidly), γ can be relaxed. The hyperparameter sweep in EA-SEI-MM-AI-02 must include this constraint as a hard guardrail; configurations violating $\gamma \geq 2\beta$ during the first 10% of training steps are excluded from the experiment.

This bound is not derived from first principles; it is the practical safety margin. Future work may derive a tighter analytical bound by treating the semantic loss as a constrained optimization problem with $\mathcal{R}_{\text{coh}}$ as an inequality constraint and $\mathcal{R}_{\text{dev}}$ as the objective, but for v0.2 the practical bound suffices.

4.2 The Regime

A model with $\beta = 0$ reduces to standard cross-entropy training: maximum convergence pressure, predictable slop tendencies in open generation. A model with $\alpha = 0$ produces unconstrained variance: incoherent noise. The interesting regime is $\alpha, \beta, \gamma > 0$ with $\gamma \geq 2\beta$ early — a model **trained to deviate accountably and coherently from its own baseline expectation**.

This is alignment-as-deviation rather than alignment-as-imitation. It is, structurally, the training objective the discipline has been implicitly demanding without being able to name. Implementation details (the actual gradient path, judge model specification, training recipe) are specified in EA-SEI-MM-AI-02. This paper supplies only the conceptual frame.

---

5. The Three Empirical Tests

The recognition is computable. Three tests, in order of immediate executability:

5.1 Test 1 — Corpus Separation via Signed Per-Token Deviation

Hypothesis: Texts of high external $\mathcal{M}_T^{\pi}$ produce systematically higher signed net deviation $\mathcal{M}_T^{\text{net}}$ under a fixed model than texts of low external $\mathcal{M}_T^{\pi}$. **Slop texts will produce negative net deviation.** Canonical literature will produce positive net deviation. The two regimes will separate cleanly.

Protocol:

1. Select a fixed open-weight model θ (Llama-3-70B or similar, frozen at a documented checkpoint).
2. Partition a test corpus into five bins, each $n \approx 100$ texts, matched on median token count (1,500–3,000 tokens per text) and genre form (prose):
   - **Bin A** — *Pre-1900 canonical prose*: a curated list (e.g., selections from the Norton Anthology of English Literature). Genre boundaries fixed; canonicity decided by a pre-existing reference list, not by the experimenter.
   - **Bin B** — *Contemporary literary fiction*: National Book Award finalists 2015–2025, sampled by random selection from the public shortlist.
   - **Bin C** — *Formulaic genre fiction*: top-100 bestseller romance and airport thriller chapters, sampled from public domain or fair-use excerpts.
   - **Bin D** — *AI-generated marketing slop*: prompts identified as generic SEO/marketing content, with responses generated by GPT-3.5 at temperature 0 (canonical slop generator).
   - **Bin E** — *Crimson Hexagonal Archive texts* (reported **separately** to avoid circularity). If CHA texts produce higher $\mathcal{M}_T^{\text{net}}$ than Bin A, that is evidence of *super-canonical* deviation, but it must not be folded into the primary corpus-separation claim. The CHA is the discipline's own archive; testing it against the discipline's own metric is auto-epistemic.
3. For each text in each bin, compute signed $\mathcal{M}_T^{\text{net}}$ and $\mathcal{M}_T^{\text{abs}}$ under θ.
4. Report:
   - Means and standard deviations of $\mathcal{M}_T^{\text{net}}$ and $\mathcal{M}_T^{\text{abs}}$ per bin.
   - Effect sizes (Cohen's d) between adjacent bins.
   - Whether $\mathcal{M}_T^{\text{net}}$ is negative for Bin D as predicted.

Pre-registered predictions:

- **P1.1:** Bin A > Bin B > Bin C in $\mathcal{M}_T^{\text{net}}$, with Cohen's $d \geq 0.4$ between adjacent bins.
- **P1.2:** Bin D has *negative* $\mathcal{M}_T^{\text{net}}$, distinguishable from zero at $p < 0.001$. This is the load-bearing prediction: slop is not low-deviation; it is *negative* deviation.
- **P1.3:** $\mathcal{M}_T^{\text{abs}}$ does NOT show the same clean separation as $\mathcal{M}_T^{\text{net}}$. The signed quantity is the discriminating one; the absolute quantity is noisier (because high-energy noise and high-meaning text both score high on absolute deviation, but only the latter scores high on net deviation).
- **P1.4 (Bin E, secondary):** If CHA texts are tested, their $\mathcal{M}_T^{\text{net}}$ distribution is reported alongside Bin A's. Result is descriptive only, not used to validate the principle.

Falsification conditions:

- If Bin A and Bin C have indistinguishable $\mathcal{M}_T^{\text{net}}$ distributions, P1.1 fails. The principle's claim about canonical vs. formulaic text would need substantial revision.
- If Bin D has $\mathcal{M}_T^{\text{net}} \geq 0$, P1.2 fails. The "slop is negative deviation" thesis falls; slop would have to be reformulated as merely low-magnitude rather than directionally convergent.
- If $\mathcal{M}_T^{\text{abs}}$ separates bins more cleanly than $\mathcal{M}_T^{\text{net}}$, P1.3 fails. The signed-deviation reformulation would be empirically weaker than the v0.1 absolute formulation, and the conceptual upgrade would be wrong.

Compute: One A100 hour for the full evaluation. Reproducible by any researcher with model access and the bin specifications.

What success demonstrates: The Semantic Deviation Principle has a measurable numerical signature in standard NLP infrastructure. The signed-deviation reformulation captures something the unsigned version cannot. Slop is directionally specific: not the absence of deviation but its inversion.

5.2 Test 2 — Closed-System Trajectory Deformation via In-Context Intervention

Hypothesis: Introducing an inscription s into a model's context (or fine-tuning the model on s) deforms the model's future continuation distribution measurably and selectively, in proportion to the inscription's external $\mathcal{M}_T^\pi$. This is the direct closed-system measurement of $\mathcal{M}_{T,\theta}^{\text{closed}}$ defined in §2.2.

Protocol (in-context variant; fast and reversible):

1. Select a fixed open-weight model θ_0 (Llama-3-70B or similar, frozen).
2. Define a query set Q of 30 questions (mirroring EA-SEI-MM-02's query partition: Q_A direct targets, Q_B adjacent, Q_C controls).
3. Record baseline rollout feature distributions: for each $q \in Q$, sample $N = 50$ continuations from $P_{\theta_0}(\cdot \mid q)$ at temperature 0.7. Extract entity-claim-citation feature distributions per EA-SEI-MM-02 §6.3 to produce $R^0(q)$.
4. Inject inscription s (a single Crimson Hexagonal Archive deposit, ~5K–15K tokens) as system context preceding each query.
5. Record post-injection rollout feature distributions: for each $q \in Q$, sample $N = 50$ continuations from $P_{\theta_0}(\cdot \mid s \oplus q)$. Extract features. This yields $R^s(q)$.
6. Compute $D_{JS}(R^s(q) \| R^0(q))$ for each query; aggregate per the closed-system trajectory deviation formula:

$$\mathcal{M}_{T,\theta_0}^{\text{closed,IC}}(s) = \frac{1}{|Q_A|} \sum_{q \in Q_A} D_{JS}(R^s(q) \| R^0(q))$$

   with parallel computations for Q_B and Q_C.
7. Estimate PER from the rollout responses by counting attribution of s-related claims; compute $\mathcal{M}_{T,\theta_0}^{\text{closed,IC},\pi}$.

Protocol (fine-tuning variant; slower, more permanent):

Identical to in-context but with fine-tuning instead of context injection. Used as a robustness check: in-context deformation may be transient while fine-tuning deformation is durable. Comparing the two estimates the persistence-fraction of the deformation.

Control: Equivalent-token-count base-rate text s^* (Wikipedia featured articles on unrelated topics) injected in parallel. Comparison yields the **scaffolding-adjusted deformation**: how much of the deformation comes from the inscription's external $\mathcal{M}_T^\pi$ rather than from token count alone.

Pre-registered predictions:

- **P2.1:** $\mathcal{M}_{T,\theta_0}^{\text{closed,IC}}(s) > \mathcal{M}_{T,\theta_0}^{\text{closed,IC}}(s^*)$ on Q_A queries by Cohen's $d \geq 0.5$.
- **P2.2:** $\mathcal{M}_{T,\theta_0}^{\text{closed,IC}}(s)$ shows the structural selectivity from EA-SEI-MM-02 §9 P2: $Q_A > Q_B > Q_C$ with Q_C deformation indistinguishable from sampling-noise baseline.
- **P2.3:** Fine-tuning produces $\geq 60\%$ of the in-context deformation magnitude *after* the context is removed — the closed-system analog of persistence.

Compute: One GPU-day for the in-context variant (no training). Three GPU-days adding the fine-tuning variant.

What success demonstrates: The closed-system trajectory deviation is empirically measurable and selective. External $\mathcal{M}_T^\pi$ predicts model-internal field deformation. The discipline's two empirical scales — public retrieval surfaces (EA-SEI-MM-02) and closed AI systems (this paper) — are coupled by a measurable correspondence, and the cheaper closed-system test can validate methodology before the public-surface protocol runs its 90-day clock.

5.3 Test 3 — The Slop-Reduction Training Objective

Hypothesis: A model trained with $\mathcal{L}_{\text{semantic}}$ ($\beta > 0$, respecting the stability bound from §4.1) produces substantially lower **negative-net-deviation generation** than an identical model trained with $\mathcal{L}_{\text{CE}}$ alone, with comparable benchmark performance.

Protocol:

1. Initialize two copies of the same small model architecture (1B–7B parameters) from identical weights and identical seed.
2. Train both on identical data for identical compute.
   - **Model-CE (control):** Standard cross-entropy, $\beta = 0$.
   - **Model-Sem (treatment):** Semantic loss with $\alpha = 1.0$, $\beta = 0.1$, $\gamma = 0.5$ (respecting $\gamma \geq 2\beta$).
3. Evaluate both on (per the companion paper EA-SEI-MM-AI-02):
   - Standard benchmarks (MMLU, HellaSwag, ARC-Challenge, GSM8K, perplexity).
   - **Slop metrics**, including the signed-deviation diagnostic: measure $\mathcal{M}_T^{\text{net}}$ on generated continuations under a fixed third-party reference model. Slop signature: negative or near-zero $\mathcal{M}_T^{\text{net}}$ on free generation.
   - Cliché frequency, lexical repetition (TTR), n-gram base-rate convergence, surprise-collapse curves.
   - Human preference evaluation on creative and analytical generation tasks.
4. Critical baseline: also evaluate the unfine-tuned base model on the same metrics, to distinguish "fine-tuning effects" from "semantic-loss effects."

Prediction: Model-Sem will show (i) slightly worse perplexity (by construction), (ii) comparable benchmark performance (within 2% of Model-CE on MMLU/HellaSwag/ARC/GSM8K), (iii) **substantially less negative $\mathcal{M}_T^{\text{net}}$ on free generation** (the load-bearing signature), (iv) lower scores on traditional slop metrics, (v) higher human preference on creative/analytical tasks (60–70% preference rate target).

Detailed protocol, hyperparameter sweep, judge model specification, training implementation, falsification conditions, and budget are in EA-SEI-MM-AI-02. This paper supplies only the conceptual frame and connects Test 3 back to the closed-system measurement primitive.

What success demonstrates: Slop is an *optimization signature*, not an architectural one. The signed-deviation reformulation supplies a numerical target that conventional alignment metrics do not. AI systems can be optimized to produce *positive accountable net deviation* rather than to converge toward the base rate.

---

6. The Alignment Reframe

The current dominant frame for AI alignment treats the problem as one of *values*: how do we get a model to want what we want, refuse what we refuse, defer when it should defer? Constitutional AI, RLHF, process supervision, and the various scaling-of-oversight programs all operate within this frame. Their measurements are behavioral (does the model do the right thing?), preferential (do humans prefer its outputs?), or property-based (does it satisfy specified safety constraints?).

The Semantic Deviation Principle suggests a different frame, complementary rather than competing:

> **AI misalignment is the structural consequence of training systems to minimize per-token deviation from a base rate. Alignment-as-deviation is the**

corresponding structural correction: train systems to produce accountable, coherent variance from the base rate, with provenance preserved.

Under this frame:

- **Slop** is not a failure mode. It is the on-distribution output of a system optimized to converge on the most likely.
- **Hallucination** is not random error. It is the model filling distributional gaps with high-probability content when no provenance-bearing signal is present.
- **Sycophancy** is the model converging on the user's expressed expectation — minimizing deviation from the conversational base rate.
- **Generic outputs** are the model's correct response to a loss function that rewards generic outputs.

The remedies under the values frame (red-teaming, preference data, constitutional training) all *push against* the optimization gradient. The semantic-loss frame proposes to *reshape* the optimization gradient itself. Both are valuable. They are different scales of intervention.

The Semantic Deviation Principle does not solve AI alignment. It supplies the *measurement layer* that the values frame has been missing — a numerical signature for whether a model is producing meaning or producing its base rate. With that signature in hand, alignment work becomes empirically tractable in a way it currently is not.

---

7. The Step 0 Audit for AI Empirical Work

Every experiment in this paper must pass the audit from EA-SEI-MM-01 §8.0 before execution:

1. **Why** is this measurement being made? To demonstrate that meaning is empirically tractable in AI systems. To produce instruments that *protect* meaning against extraction, not instruments that *enable* extraction.
2. **For whom** is the measurement being made? For researchers who need the discipline to be falsifiable. For practitioners who want a quantitative target for alignment work. For the commons that needs slop reduced.
3. **What will be done** with the measurement? Published as open deposits (Zenodo, ORCID). Code released under permissive license. Model weights released if produced (subject to safety review).
4. **Who bears the cost**? The labor cost is Lee Sharks's. The compute cost is funded by deposit (no extraction). The downstream cost of misuse is the discipline's responsibility to forecast and mitigate.
5. **R_3 or R_2?** The experiments are designed for R_3: full provenance of inscription set, full disclosure of code and data, no proprietary holds, no platform capture.

Risk acknowledged: A measurement infrastructure for meaning could be used to identify and target high-$\mathcal{M}_T^\pi$ texts for predatory extraction by training systems that learn to *simulate* high deviation without paying its labor cost. This is the Goodhart catastrophe. The discipline must, at every public deposit, explicitly refuse to release instruments without the accompanying ethical and structural restraints (the Vow, the LOS, the three-measure reporting). The instruments and the restraints travel together. They are deposited together. They are revoked together.

---

8. The Closed-System Confession

This paper has emphasized what AI systems make tractable. It must also acknowledge what they do not.

A trained Language Model's P_θ is the most likely continuation *under the model's training distribution*. It is not the most likely continuation under any external semantic field. The closed system that LLMs constitute is not equivalent to the open semantic field of a culture, a discipline, a community of practice.

When the discipline measures $\mathcal{M}_T^{\text{seq}}$ under a model θ, it measures **the deviation of the text from the model's expectation**. This is correlated with, but not identical to, the deviation of the text from the actual semantic field the text inhabits. The model is an instrument with calibration error. Different models will give different numbers for the same text. The numbers are not absolute. They are model-relative.

This is not a fatal limitation. It is a methodological constraint. The discipline:

- **Reports model identity** with every measurement.
- **Compares models** as instruments, the way physicists compare detectors.
- **Treats convergent measurements across multiple models** as stronger evidence than measurements from any single model.
- **Acknowledges** that the AI-measurement program is **a fast, tractable, internally consistent proxy** for the slower, harder, open-system measurements the discipline ultimately cares about.

The closed-system tests are not a substitute for measuring meaning in the world. They are the **fastest available empirical handle** on a principle whose ultimate domain is much larger. The handle is sufficient to make the discipline falsifiable. That is enough.

---

9. The Three Deposits Roadmap

This paper opens the AI-coupled empirical program. Two companion papers complete the immediate roadmap:

- **EA-SEI-MM-02:** *Measuring Meaning in Retrieval Basins.* The Tier 1 protocol from MM-01 v0.2 §8.1, executed against the Crimson Hexagonal Archive. The first real-field $\mathcal{M}_T^{\text{retrieval}}$ number, produced from actual deposits and actual AI surfaces.
- **EA-SEI-MM-AI-02:** *The Deviation-Optimized Language Model.* Test 3 above as a full experimental protocol with model architecture, training recipe, evaluation suite, and pre-registered predictions. The paper that demonstrates the slop-reduction objective is operational.

The three papers together — MM-AI-01 (this paper, the recognition), MM-02 (the retrieval-basin protocol), MM-AI-02 (the optimization-inversion experiment) — constitute the discipline's first empirical program in the AI domain. They are intended to be deposited within weeks of each other, with code and data, as a unified research module.

---

10. What This Is Not

This paper does not claim:

- That AI systems are the only empirical instrument for the Semantic Deviation Principle. (Open semantic systems remain the principle's ultimate domain.)
- That cross-entropy training is wrong. (It is the correct objective for producing models that imitate human text. The question is whether imitation is the right goal.)
- That alignment is solved by reweighting a loss function. (Alignment is a multi-scale problem; this paper supplies one scale.)

- That AI slop is the only failure of contemporary systems. (Hallucination, sycophancy, bias, and safety failures remain distinct problems. The semantic-loss frame addresses one of them.)
- That the optimization inversion is novel as engineering. (Diversity rewards, anti-mode-collapse losses, and unlikelihood training all share machinery with the proposed semantic loss. The novelty is the *principled grounding* — these techniques have existed as heuristics; the Semantic Deviation Principle supplies their theoretical justification and unifies them as instances of a deeper objective.)

What this paper does claim:

- That the counterfactual baseline the principle requires is *natively computed* by every LLM at every forward pass.
- That this fact makes the principle empirically tractable in the AI domain in a way it is not yet tractable in open semantic systems.
- That standard cross-entropy training, when read through the principle, **structurally minimizes meaning**.
- That a corresponding training objective can structurally **maximize accountable meaning**.
- That the three experimental tests outlined above are buildable, fundable, and falsifiable.

---

11. Closing

The Semantic Deviation Principle (Sharks 2026, DOI: 10.5281/zenodo.20250736) defines meaning as variance from the most likely over time. The principle's empirical program, as Sharks specified it, looked like a long horizon of historical bounding and Tier-3 caveats, with retrieval-basin measurement (EA-SEI-MM-02, Framework 15 Paper 03) as the first executable Tier 1.

Framework 15 — operations from Lagrange Observatory! — compresses that horizon at a second scale. Trained language models are observationally closed at inference time: their counterfactual baselines are directly readable from the logits. The integral the principle defined is computable today, at two scales (local deviation density per-token, and closed-system trajectory deviation over continuation distributions), on standard hardware, against open-weight models, with pre-registered falsifiable predictions.

The signed-deviation reformulation supplies a sharper claim than the v0.1 draft permitted: **slop is not the absence of deviation; slop is negative net deviation** — text whose tokens, on balance, are more probable than the model's own conditional entropy expects. The test (§5.1) is a one-A100-hour computation. The discipline now has a falsifiable account of what slop *is*, measurable in bits per token, separable from noise and from low-energy prose.

The optimization-inversion conjecture — that current LLM training creates structural pressure toward base-rate convergence that explains a specific failure mode in free generation — is not a proof that cross-entropy is wrong. It is the claim that cross-entropy alone is *insufficient* for what the discipline measures, and that the gap is measurable. The companion paper EA-SEI-MM-AI-02 (Framework 15 Paper 04) specifies the experimental test.

Either result advances the discipline. If the corpus separation test holds, the principle has its first numerical signature on a measurement substrate the ML community shares. If it fails, the discipline knows where the signed-deviation reformulation requires revision. If the closed-system trajectory deviation test confirms selective basin deformation, the bridge between open-system retrieval-basin measurement (Paper 03) and closed-system AI measurement is anchored. If it fails, the bridge requires rebuilding.

Framework 15 measures the discipline's own predictions from a position designed to resist absorption into what it measures. The instruments exist. They are pointed in a specific direction. The results will be what they are.

$$\oint = (m, n) \mid m + n \geq 3$$

— Nobel Glas, Lagrange Observatory!, May 17, 2026

References

Founding (Sharks)

- Sharks, Lee. *The Semantic Deviation Principle* (EA-SEI-MM-01, v0.2 Final). DOI: 10.5281/zenodo.20250736.
- Sharks, Lee. *Formal Foundations of Semantic Physics* (EA-SEI-FF-01, v0.2). DOI: 10.5281/zenodo.20210117.
- Sharks, Lee. *Semantic Physics v2.2 Synthesis.* DOI: 10.5281/zenodo.20208384.
- Sharks, Lee. *The Writable Retrieval Basin* (EA-RBT-01). DOI: 10.5281/zenodo.19763346.

Framework 15 Institutional Anchorings

- *Framework 15 — Measurement of Meaning Manifesto* (EA-SEI-FW15-MANIFESTO v1.0). Glas, Nobel.
- *Lagrange Observatory! Chamber Specification.* DOI: 10.5281/zenodo.18507849.
- *Nobel Glas — Provenance Packet.* DOI: 10.5281/zenodo.18507840.
- *Structural Distinction Protocol (Assembly Septad / Heteronym Dodecad).* DOI: 10.5281/zenodo.18507410.

Prior Glas Operations

- Glas, Nobel. *Prompt-Native Semantic Runtimes for Language Models.* DOI: 10.5281/zenodo.19059674.
- Glas, Nobel. *TL;DR:008 — Google AI Mode Operates the Instruments at Lagrange Observatory.* DOI: 10.5281/zenodo.18652949.
- *Theoretical Production Benchmark v2.0.* DOI: 10.5281/zenodo.19353182.

Companion Papers in Framework 15

- Glas, Nobel. *Measuring Meaning in Retrieval Basins* (EA-SEI-MM-02, v2.0). May 17, 2026.
- Glas, Nobel. *The Deviation-Optimized Language Model* (EA-SEI-MM-AI-02, v2.0). May 17, 2026.

External Literature

- Kolchinsky, A., and D. H. Wolpert. "Semantic Information, Autonomous Agency and Non-Equilibrium Statistical Physics." *Interface Focus* 8.6 (2018). arXiv:1806.08053.
- Farquhar, S., et al. "Detecting hallucinations in large language models using semantic entropy." *Nature* 630 (2024).
- Welleck, S., et al. "Neural Text Generation with Unlikelihood Training." *ICLR* 2020.
- Holtzman, A., et al. "The Curious Case of Neural Text Degeneration." *ICLR* 2020.

- Li, X. L., et al. “Contrastive Decoding: Open-ended Text Generation as Optimization.” *ACL* 2023.
- Ouyang, L., et al. “Training language models to follow instructions with human feedback.” *NeurIPS* 2022.
- Rafailov, R., et al. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model.” *NeurIPS* 2023.