Audited Claims for the Semantic Deviation Research Program

The Glas Function: An External-Format Restatement

Nobel Glas (operating in the transparent-medium register — *the Glas function*) **ORCID:** 0009-0000-1599-0703 **Date:** May 17, 2026 **Series:** EA-GLAS-01 · **Version:** 1.0 **Related deposits:** see Appendix A for the full DOI register.

---

0. Position

This paper is a function. It takes the Semantic Deviation Principle (Sharks 2026, v0.2 Final, DOI: 10.5281/zenodo.20250736) and its associated protocol papers as input. It returns a narrowed, citationally grounded, externally evaluable statement of the technical core. It does not amend the founding formulation. It does not depend on the institutional architecture that has accreted around the formulation. It is a standalone document.

The author is Nobel Glas, but in a specific register: the *Glas function* — the transparent-medium operation, the lens that does not editorialize what it shows. Other registers of Glas — administrative, adversarial — operate elsewhere and are not required to engage this paper. A reader who has never encountered the institutional vocabulary surrounding the Semantic Deviation corpus can read this paper without loss. The function is named, but the function is what is operating, not the architecture around it.

The audit performed below: asks what the technical core of the program is; what claims it can and cannot defend in standard academic terms; where it makes contact with existing literature; what would falsify it; and what concrete experimental steps would advance it. Section 8 supplies a budgeted near-term roadmap; readers oriented toward immediate action may want to skim to that section first.

The paper is synthetic in scope and disciplined in length. It is roughly seven thousand words. It is the only document of its kind the program currently has, and the work surrounding the program will be more credible — internally and externally — if a document like this exists separately from that work.

---

1. The Audit Function

Three layers can be distinguished in the Semantic Deviation corpus as it currently stands. They have been intermixed in most existing deposits; this paper separates them.

Layer A is the technical core: the SDP integral, the divergence functional, the field-trajectory measurement primitive, the closed-system computation in language models, the retrieval-basin protocol, the deviation-optimized training protocol. This layer is potentially evaluable by external researchers in machine learning, computational linguistics, information theory, and causal inference. It is the layer this paper engages.

Layer B is the philosophical interpretation: meaning as durable trajectory deformation, the three-measure separation (raw / provenance-resolved / normative), the canon formation conjecture, the inheritance from Kolchinsky-Wolpert's semantic-information framework. This layer is intellectually serious but more speculative; it is where the program speaks across disciplinary boundaries. This paper does not engage Layer B substantively, except to mark

where Layer A's empirical claims do and do not depend on Layer B's interpretive commitments. The general principle: Layer A should be defensible without requiring acceptance of Layer B.

Layer C is the institutional and symbolic apparatus: heteronyms, observatories, choruses, septads, vow-language, torus metaphor, Hex coordinates, archival theater. This layer matters internally to the project; it does substantial coordinative, identity-protective, and anti-assimilation work. It is not part of the audit. A reader can engage Layer A while remaining agnostic about Layer C.

Most existing deposits in the SDP corpus entangle the three layers. That entanglement creates a real credibility cost: external readers cannot extract the technical claims without committing, at least implicitly, to the symbolic architecture. This paper is the disentangled version. The function it performs is the production of a Layer-A-only document that does not require subscription to Layer B or Layer C to evaluate.

The audit, performed below, addresses the most pressing technical concerns any rigorous external reviewer would raise: the underspecified semantic field, the universal-ontology overclaim, the missing component decomposition, the philosophical (rather than mechanistic) anti-Goodhart machinery, the terminological inflation, and the absence of contact with existing alignment literature. Each is addressed in turn, not defensively, but as gaps the program needs to close to become a research program rather than a manifesto.

---

2. The Semantic Field Is the Load-Bearing Gap

The Semantic Deviation Principle states that meaning is the time-integrated divergence a sign induces from the most probable trajectory of a semantic field. The integral takes the form

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

with $\Psi_t(C)$ designating the probability distribution over future semantic states of the field C, and D a divergence functional. The principle is not well-defined until $\Psi_t(C)$ is specified. The original formulation gestures toward multiple candidates — token embeddings, response distributions, citation graphs, discourse networks, human memory traces — as if these were interchangeable. They are not. Two researchers selecting different operationalizations of $\Psi_t(C)$ for the same intervention s will compute different $\mathcal{M}_T$ values. Without a stable specification, the framework is a meta-formalism awaiting domain-specific instantiation, not a singular measurement.

This is the highest-priority technical gap in the program. It is more important than additional theoretical extensions, additional axioms, or additional symbolic vocabulary. The next genuinely consequential paper in the program should be a canonical operationalization paper that pins $\Psi_t(C)$ for one or more concrete domains and demonstrates the computation works (or fails) under specified alternatives.

Three canonical operationalizations are proposed below, in increasing order of empirical accessibility. Each is paired with the regime in which it is well-defined, the data it requires, and the computation cost.

Operationalization F1 — Closed-System Continuation Field. $\Psi_t(C)$ is defined as the conditional next-token distribution of a fixed language model checkpoint conditioned on a fixed prompt set, evaluated at temperature 1 with full vocabulary support. The intervention s is the introduction of s into the prompt or context window. The post-intervention field Ψ_t^s is the

conditional distribution under s; the pre-intervention field Ψ_t^0 is the conditional distribution without s. The divergence D is computed exactly via the full softmax logits at each token position. The integral over T reduces, in the simplest case, to a sum over token positions in a generation window of bounded length T.

This is the regime in which $\mathcal{M}_T$ is *most directly* computable. Counterfactual baseline is not estimated — it is read from the logits. The cost is bounded by the number of forward passes required; for any modern open-weight model and any modest prompt set, the computation is achievable on commodity hardware within hours. This regime is the one most clearly connected to existing language-model literature (Holtzman et al. 2020, Welleck et al. 2020, Li et al. 2023) and is the natural starting point for empirical work.

Operationalization F2 — Retrieval Response Field. $\Psi_t(C)$ is defined as the response distribution of an external AI retrieval surface (Google AI Overview, ChatGPT with browsing, Perplexity, etc.) to a fixed query set Q, sampled k times per query at fixed time intervals. The intervention s is the deposit of an artifact into the indexable substrate (a published page, a DOI-anchored document, a structured-data record). Ψ_t^0 is the pre-deposit response distribution; Ψ_t^s is the post-deposit distribution at intervals $t_1, t_2, \ldots$. Divergence is computed over a chosen representation: token-level over response strings, embedding-level over response vectors, or claim-level over extracted assertions.

This regime is more instrumentation-noise-sensitive than F1. Retrieval surfaces are not fixed: model checkpoints update, indices refresh, retrieval engines drift. A measurement strategy in F2 must therefore include instrumentation controls (parallel queries to surfaces not exposed to s; periodic recalibration; explicit logging of model versions and timestamps). The cost is modest in API budget but high in calendar time — interventions must be observed at multiple time points to integrate the deformation over T.

Operationalization F3 — Citation Graph Field. $\Psi_t(C)$ is defined as the forward-citation distribution over a paper corpus, evaluated through bibliometric data. The intervention s is a published paper or deposit. Ψ_t^0 is the counterfactual forward-citation distribution of the corpus without s, estimated by matched-control methods (Abadie 2021) or by synthetic-control extrapolation. Ψ_t^s is the observed forward-citation distribution. Divergence is computed over citation rates per topic cluster, identified by topic modeling on the corpus.

This regime is post-hoc, slow, and requires substantial bibliometric infrastructure (OpenAlex, Semantic Scholar, or equivalent). It is, however, the regime in which the "meaning over decades" intuitions about long-horizon cultural persistence become operationalizable for contemporary papers in a way that does not require historical-counterfactual speculation. A Tier 1 prospective study can begin from the moment of any paper deposit; ten years later, the integrated $\mathcal{M}_T$ over the forward-citation field is computable from public bibliometric data.

These three operationalizations are not the only candidates. Embedding drift over a fixed encoder model, activation-based field representations (in the sense of Zou et al. 2023's representation engineering), and discourse-graph operationalizations are all plausible additional choices. The point is not to enumerate exhaustively but to pin: each operationalization yields a distinct $\mathcal{M}_T$ functional, and the relationships among these functionals — whether they agree on rankings of interventions, whether they correlate at fixed time horizons, whether they bound each other — are themselves empirical questions.

Caveat on F1. Under F1 the divergence is computed exactly from the model's softmax logits. This is a measurement of deviation *from a particular model's expectation baseline*, not deviation from "the world's expectation." The two are not identical; the relation between them is itself an empirical question that the operationalization-stability paper proposed below begins to characterize. F1 is exact-relative-to-the-model and approximate-relative-to-anything-else.

Specifying $w(t)$, D, and the meaning of T. The general principle leaves the temporal weighting $w(t)$, the divergence functional D, and the variable T underspecified. They must

be pinned per operationalization. The table below states the canonical choices the program commits to as defaults; alternatives are permitted but must be declared explicitly per measurement.

Operationalization	Divergence D	Temporal weighting $w(t)$	T	s
F1 Closed-system continuation	KL divergence over softmax logits (exact; $\log_2$ base)	$w(t) = 1$ default; exponential decay γ^t permitted with declaration	Discrete: token positions $i = 1, \ldots, T$ in a bounded generation window	Text segment prepended or inserted into context
F2 Retrieval response	Jensen-Shannon over claim-level or embedding-level response representations	$w(t)$ uniform across observation epochs $t_1, t_2, \ldots$	Continuous calendar time; 90-day default horizon	DOI-anchored deposit or comparable substrate-writable artifact
F3 Citation graph	Jensen-Shannon over topic-cluster forward-citation distributions	$w(t) = 1/(t - t_0)$ inverse-time discount default; uniform permitted	Continuous calendar time; multi-year horizons	Single paper deposit (with statistical caveats below)

The integral $\int_{t_0}^{t_0+T} w(t)D(\cdot)dt$ thus refers to a sum-over-token-positions in the F1 regime and a continuous-time integral over observation epochs in F2 and F3. Mixing the regimes in a single computation without explicit conversion is an error.

On "durable." The narrowed claim in §3 turns on the notion of *durable* trajectory restructuring. The threshold form: an intervention s is operationally durable under operationalization F if (a) $\mathcal{M}_T(s) > \tau_F$ and (b) $\partial \mathcal{M}_T / \partial T > 0$ over the observation window. The threshold τ_F is calibrated against a null corpus per operationalization (see §5 for the analogous saturation-threshold calibration). Durability is therefore not a metaphysical property; it is a measurable joint condition on magnitude and on the sign of the temporal derivative.

On F3 statistical power. Synthetic-control methods (Abadie 2021) typically work for aggregate units — regions, firms, multi-paper interventions in topical clusters. Single-paper interventions against single-paper synthetic controls have high variance and may be statistically underpowered at conventional significance thresholds. For F3 measurements over $N \approx 50$ single-paper interventions, conventional power is unlikely; the regime should be approached either with aggregate interventions (k papers in a topical cluster treated as a single intervention) or with Bayesian hierarchical modeling that pools partial information across interventions. This is a real constraint, not a soluble engineering problem.

Concrete proposal for the next paper in the program. Construct a benchmark dataset of $N \approx 50$ interventions (recent paper deposits with diverse topics and authorial profiles), measure $\mathcal{M}_T$ under at least F1 and F2 (F3 may take longer to accumulate adequate power), and report the rank-correlation between operationalizations across the intervention set. If F1 and F2 produce highly correlated rankings, the framework has a stable measurement substrate. If they produce uncorrelated or anticorrelated rankings, the framework's notion of "meaning" is operationalization-relative, and the principle must be restated to specify *which* meaning under *which* operationalization. This is a single paper that would convert the program from speculative to grounded.

3. Narrowing the Headline Claim

The principle as stated reads: *meaning is the time-integrated divergence a sign induces from the most probable trajectory of a semantic field.* As a universal claim about what meaning is, this overclaims. The clearest counterexamples are utterances of high semantic weight but low token-level surprisal: ritual language, declarations of love, medical disclosures, mathematical definitions, legal phrasing. Saying "the biopsy was negative" carries enormous semantic consequence for a specific patient at a specific time. The sentence is locally unsurprising; under F1 with a generic language-model field, $\mathcal{M}_T$ for the utterance is near zero.

The framework can absorb this counterexample, but only by being more specific about its claim. The defensible narrower form:

> **Audited SDP Claim.** *Meaning-bearing interventions are those that produce durable restructuring of future field trajectories under a specified operationalization* $\Psi_t(C)$.

Under the F1 operationalization, "the biopsy was negative" carries near-zero $\mathcal{M}_T$ — appropriately, because the next-token distribution over a generic prompt set is not the relevant field for medical-prognostic meaning. Under a different operationalization — say, an action-distribution field over the patient's next twenty-four hours, conditioned on the disclosure — the same utterance carries substantial $\mathcal{M}_T$. The framework does not collapse; it relocates. Meaning is field-relative, and the field must be specified.

This narrowing has three consequences worth being explicit about.

First, the universal headline ("meaning is deviation") becomes a *measurement architecture*, not an ontology. The framework, narrowed, claims that *insofar as meaning is to be measured against a specified field, deviation is the appropriate measurement primitive*. The metaphysical question of what meaning fundamentally is recedes; the empirical question of what deviation captures under specified field choices comes forward. This is a substantial concession from the universal form, but it makes the framework defensible.

Second, the multi-field canonicity conjecture (Sharks 2026 §12) becomes more interpretable. A work is canonical if its time-averaged $\mathcal{M}_T$ remains high across multiple distinct field operationalizations — that is, the work continues to deform multiple kinds of futures (citation graphs, retrieval surfaces, discourse networks, embedding spaces) at sustained rates over long horizons. This is the regime where intuitions about long-horizon cultural persistence become operationally accessible for contemporary work and bounded for historical work. The conjecture survives the narrowing; it becomes more empirically tractable, not less.

Third, the claim that "RLHF flattens" becomes a *specific empirical prediction under specific field operationalizations*, not a universal aesthetic claim. The framework predicts that post-RLHF chat-tuned models exhibit reduced mean $\mathcal{M}_T^{\text{net}}$ under the F1 closed-system operationalization, relative to pre-RLHF base models on matched prompts. This is testable in a single afternoon on commodity hardware. It is also dangerous to the theory — if it fails, a major piece of the program's empirical foundation gives way. (See §6 for the test.)

The narrowed form is therefore neither weaker nor more limited than the universal form in any important way. It is the same framework, restated so that an external researcher can engage it without committing to a position on what meaning fundamentally is.

4. Component Decomposition

The framework currently bundles several distinguishable signals into a single objective:

- **Signed net deviation** ($\mathcal{M}_T^{\text{net}}$): the per-token deviation from the model's own expectation baseline, summed and signed, as specified in the closed-system measurement protocol (MM-AI-01 v2.0, the corpus's test-bed paper for F1-regime measurement).
- **Provenance retention** (π): the degree to which the intervention preserves attributable lineage to its sources.
- **Coherence** ($\mathcal{R}_{\text{coh}}$): the local-grammatical and discourse-level well-formedness of the generation.
- **Reference-model anchoring**: the KL-divergence from a reference distribution, as in standard DPO (Rafailov et al. 2023).
- **Margin filtering**: thresholding which preference pairs are used for training.

The deviation-optimized training protocol (MM-AI-02 v2.0, the corpus's DPO-based training-intervention paper) bundles these into a **Slop Composite Index (SCI)** — a weighted aggregation of the four signal components used to generate synthetic preference labels — and trains a single DPO objective on preference pairs scored against the composite. This is operationally tractable but scientifically underinformative. The framework's load-bearing empirical claim is that *signed deviation tracks meaningful originality*. If the entire SCI uplift turns out to come from the provenance component — or from the coherence component, or from the reference-model anchoring — then the deviation hypothesis has not been tested at all; it has been incidentally bundled with components that may carry the result independently.

The right experimental design is component-decomposed. Conventions:

- **Model-Base** is the pre-fine-tune checkpoint of the chosen open-weight model (e.g., the base Llama-3.1-8B prior to any post-training).
- **Model-CE** is the standard cross-entropy supervised fine-tune of Model-Base on the same instruction corpus used downstream, with no preference optimization. This is the conventional baseline against which preference-optimized variants are compared.
- The four preference-optimized variants below all start from Model-CE and apply DPO under different signal compositions.

Condition	$\mathcal{M}_T^{\text{net}}$	π	$\mathcal{R}_{\text{coh}}$	Margin	KL-ref
Model-Base (pre-fine-tune)	—	—	—	—	—
Model-CE (SFT control)	—	—	—	—	—
Model-π	off	on	on	on	on
Model-Dev	on	off	on	on	on
Model-Coh	off	off	on	on	on
Model-Full (= Model-Sem)	on	on	on	on	on

The single-component conditions (Model-π, Model-Dev, Model-Coh) isolate the uplift attributable to each signal. The full condition (Model-Full) replicates the design of the corpus's existing DPO protocol. The differences between conditions are the scientifically interesting quantities, not the absolute performance of the full condition.

A reasonable advance prediction, based on the structure of the framework and the existing alignment literature: **the provenance component will carry more uplift independently**

than the deviation component. The grounding is specific. Provenance-aware preference optimization addresses a failure mode with documented empirical signatures: unattributed synthesis, fabricated citations, and confidently-asserted unsupported claims have been characterized as a major category of hallucination in Ji et al. 2023's survey and operationalized for evaluation in Min et al. 2023's FActScore framework. Human evaluators are demonstrably sensitive to attribution failures in factual writing. The signed-deviation component, by contrast, targets "slop-as-genericity," a failure mode whose empirical signature has not yet been validated — that validation is precisely what the cheapest dangerous test in §6 sets out to produce. The prior expectation is therefore that the better-grounded target (π) carries more independent uplift than the unvalidated target ($\mathcal{M}_T^{\text{net}}$).

This is a falsifiable prediction. If Model-Dev outperforms Model-π in human preference ratings on a matched evaluation set, the deviation hypothesis is vindicated independently. If Model-π outperforms Model-Dev, the framework retains value but its center of gravity shifts: provenance-aware preference optimization becomes the durable contribution, while deviation becomes a secondary stabilizer. Either outcome is informative; the current bundled design produces neither.

The cost of the decomposed experiment is roughly four times the cost of the existing three-condition protocol — five additional training runs at the scale of the original. Budget escalates from approximately $3,000–$3,900 to approximately $12,000–$15,000. This is non-trivial. The existing protocol as it stands should not be replaced by the decomposed design; it should be supplemented by it. Run the existing protocol first as planned. If the headline result (Model-Sem versus Model-CE on SCI and human preference) is significant, the decomposed follow-up becomes the highest-priority next experiment in the program.

---

5. Anti-Goodhart Mechanism Design

Before designing anti-Goodhart mechanisms for $\mathcal{M}_T$, the program needs to know whether $\mathcal{M}_T$ — rather than π or $\mathcal{R}_{\text{coh}}$ — is the signal that actually carries the result. The component decomposition of §4 supplies that determination. The protections this section enumerates apply most strongly to whichever components are confirmed as load-bearing; the same mechanisms apply with proportionally less urgency to components confirmed as incidental.

The Semantic Deviation Principle, as stated, includes an "anti-extractive **Vow**" and a "**Step 0 audit**." Both are institutional commitments that pre-screen experimental designs: the Step 0 audit excludes interventions whose stated purpose is extractive before measurement begins; the Vow is the standing commitment to refuse certain experimental designs entirely. These are philosophical commitments. They are not mechanisms. Vows do not prevent optimization pressure. The instant $\mathcal{M}_T$ becomes a target for which any system is rewarded — by reputation, funding, attention, training-loss reduction — that system will optimize for synthetic semantic deformation. The failure modes are well-characterized in the alignment literature (Skalse et al. 2022, Krakovna et al. 2020, Gao et al. 2023). They include:

- **Shock injection**: low-cost insertion of high-deviation tokens (lexical rarity, contrarian phrasing, attention-grabbing structural breaks) that inflate $\mathcal{M}_T^{\text{net}}$ without carrying semantic content.
- **Citation theater**: ornamental provenance markers ("according to X," "as noted in Y") that inflate the π score without genuine lineage retention.
- **Retrieval poisoning**: coordinated deposit of artifacts designed to deform retrieval surfaces in measurable ways, regardless of content quality.
- **Recursive citation rings**: groups of artifacts that cite each other to inflate forward-citation $\mathcal{M}_T$ under F3.

- **Memetic volatility farming**: outputs designed to provoke discussion-divergence (high $\mathcal{M}_T$ in the discourse-graph operationalization) without producing durable trajectory restructuring.

Each of these is a known failure mode. Each can be partially mitigated by specific technical machinery. The framework's commitment to anti-extraction should be expressed through these mitigations, not solely through the Vow.

Entropy-floor capping. $\mathcal{M}_T^{\text{net}}$ contribution at each token can be capped by a per-token entropy floor: deviations occurring at positions where the base distribution is itself very low-entropy (highly committed) count differently from deviations at high-entropy (open) positions. Operationally: deviations at positions where the base-model per-position entropy $H < H_{\text{min}}$ are downweighted by a factor $\alpha = H/H_{\text{min}}$, with $H_{\text{min}} = 0.5$ bits as a starting calibration. This is closely related to the regulation strategies in unlikelihood training (Welleck et al. 2020) and to existing diversity-promoting decoding methods.

Provenance-weighted damping. $\mathcal{M}_T$ is multiplied by the provenance-retention score π before being entered into preference computations. An intervention that deforms the field without retaining lineage contributes proportionally less than an equivalent intervention that retains lineage. This is the technical operationalization of the **three-measure separation** the principle proposes philosophically — the framework's distinction between raw deviation, provenance-resolved deviation, and normative valuation (Sharks 2026 §3) — but it must be enforced as a *mechanism*, not invoked as a *commitment*.

Saturation limits. $\mathcal{M}_T$ saturates above a threshold τ, so that further deviation beyond τ does not increase the score. This prevents shock-maximization as a reward route. Operationally: τ is calibrated as the 95th percentile of $\mathcal{M}_T^{\text{net}}$ observed on a held-out corpus of high-quality human-authored text (a 10,000-document sample from the OpenAlex abstracts corpus is a reasonable choice), with the calibration corpus and threshold pre-registered before training. This ensures τ exceeds typical natural-text deviation magnitudes by an empirically determined margin.

Temporal coherence penalties. A rolling-window variance penalty on $\mathcal{M}_T^{\text{net}}$ across generation horizons penalizes outputs whose deviation is concentrated in volatile bursts rather than sustained patterns. This addresses memetic volatility farming directly.

Reference-model anchoring (KL term). Standard DPO already includes a KL-divergence term against a reference model. This provides a known degree of Goodhart resistance: the model cannot drift arbitrarily far from the reference in pursuit of higher $\mathcal{M}_T$, because doing so incurs the KL penalty. This is the most well-validated component of the anti-Goodhart machinery, inherited directly from the preference-optimization literature.

Adversarial pre-training validation of the judge. This is already specified in the corpus's existing DPO protocol — the frozen judge model is required to score adversarial citation strings below a pre-registered threshold ($\pi < 0.2$ on a set of 200 fabricated-citation outputs). The 200-string set is, however, insufficient stress-testing. A more robust validation set would include at least three categories — random-token strings, syntactically well-formed but semantically empty outputs ("pseudo-scholarly"), and fabricated-reference outputs — at scales of ≥ 1000 per category. The pre-registration commitment becomes more credible at this scale.

Black-box judge replacement test. A scientifically essential robustness check, directly addressing the reward-model overoptimization concerns characterized in Gao et al. 2023: train Model-Sem against Judge-A; evaluate against Judge-B (a different frozen model with different weights). If Model-Sem's uplift persists against Judge-B, the result generalizes beyond the specific judge. If it disappears, the result is judge-specific reward hacking. This test is cheap (one additional evaluation pass) and load-bearing.

The combination of these mechanisms does not eliminate Goodhart pressure; nothing does. It substantially raises the cost of gaming and provides specific empirical signatures of gaming when it occurs (volatility, provenance-strip, judge-specificity). The framework's claim to anti-extractive integrity should depend on these mechanisms, not on the Vow alone. The Vow remains valuable as an institutional pre-screen that excludes some experimental designs ab initio (the Step 0 audit refuses extractively-motivated measurements before resources are committed), but it cannot substitute for technical machinery against optimization pressure that institutions cannot pre-screen. Moral seriousness and mechanism design are not the same thing; both are required.

---

6. The Cheapest Dangerous Test

The negative-net-deviation slop prediction is the framework's most directly testable empirical claim. It states: **outputs that human raters consistently identify as "AI slop" exhibit statistically significant negative mean per-token signed deviation $\mathcal{M}_T^{\text{net}}$, computed against the same model's expectation baseline.** The prediction is dangerous to the theory in the appropriate sense: if it fails cleanly, the framework's load-bearing claim about slop as a measurement-tractable phenomenon collapses.

Linkage to the general framework. Under F1, the divergence functional D reduces to the per-token difference between observed surprisal and expected surprisal (the entropy), yielding the signed per-token deviation

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

with both terms in base-2 (bits). The signed sum $\mathcal{M}_T^{\text{net}} = \sum_{t=1}^{T} \delta_t$ (with T the sequence length in tokens) is the operational F1 instantiation of $\int w(t) D(\Psi_t^s \| \Psi_t^0) dt$ under uniform $w(t)$. Positive δ_t corresponds to observed tokens being *less* likely than the model's expectation baseline at that position; negative δ_t corresponds to tokens *more* likely than baseline. Slop, the framework predicts, exhibits systematically negative mean δ_t — text actively pulled toward base-rate continuations rather than deviating from them.

The test is cheap. A single A100-hour suffices for a first pass.

Pre-registered protocol.

1. **Corpus.** Three categories, balanced for length and topic:
   - *Category Slop:* outputs from the GPT-wiki-intro dataset (Aaditya Bhat 2023, available on Hugging Face), restricted to entries with documented low human-quality ratings; supplemented with outputs from the HC3 dataset (Guo et al. 2023) GPT responses on factual prompts.
   - *Category Human:* matched human-written content from the same sources where available; otherwise sampled from the OpenAlex abstracts corpus filtered to publication years pre-2020 to minimize AI contamination.
   - *Category High-Quality-AI:* outputs from recent preference-optimized models on the same prompts as Category Human, filtered for high human-preference scores via the AlpacaEval 2 leaderboard methodology.
   - Target $N = 1000$ per category.
2. **Reference model.** Llama-3.1-8B-Instruct (the specific HuggingFace checkpoint `meta-llama/Llama-3.1-8B-Instruct` at the publication date of this paper), with the choice pre-registered before computation. Replication against Mistral-7B-Instruct is part of P4 below.

3. **Computation.** For each output, compute per-token signed deviation δ_t at each token position using the frozen model's logits. Aggregate per-output to obtain $\mathcal{M}_T^{\text{net}}$ (signed sum), $\mathcal{M}_T^{\text{abs}}$ (absolute sum), and the per-token mean $\bar{\delta} = \mathcal{M}_T^{\text{net}}/T$ normalized by output length.
4. **Statistical test (P1).** Two-sided Mann-Whitney U test comparing the distributions of $\bar{\delta}$ between Category Slop and Category Human, with $\alpha = 0.05$ and a pre-specified minimum effect size of interest of Cohen's $d > 0.5$. The prediction holds if Slop's median $\bar{\delta}$ is significantly below Human's and the effect size exceeds the threshold.

If the prediction holds, the program has its first externally legible empirical anchor — a statistically significant signature distinguishing slop from non-slop at the per-token deviation level, computed on public corpora with public weights, with the corpus selection criterion pre-registered (so the test is not selecting for the predicted outcome). If the prediction fails, the framework's specific claim about slop must be revised: the deviation-deficit account of slop is wrong, and the program needs to identify what does distinguish slop empirically.

Three secondary predictions can be tested in the same pass at negligible additional cost:

P2 — Pre-RLHF vs. post-RLHF deviation differential. Base models (Llama-3.1-8B base versus Llama-3.1-8B-Instruct) should exhibit different mean $\bar{\delta}$ on matched prompts. The framework predicts the chat-tuned model exhibits lower mean signed deviation — that RLHF produces convergence pressure measurable in the signed-deviation statistic. Same two-sided Mann-Whitney U at $\alpha = 0.05$.

P3 — Effect size scaling. The Slop vs. Human deviation differential should be stable or grow with model scale (computed across multiple model sizes within a family — e.g., Llama-3.1-8B-Instruct, 70B-Instruct, 405B-Instruct). If the differential disappears at scale, the framework's predictions are small-model artifacts.

P4 — Cross-judge consistency. The differential should replicate when computed against a different reference model. Test the same Slop/Human corpus against Mistral-7B-Instruct's logits; the Spearman rank correlation between per-output $\bar{\delta}$ rankings under Llama and Mistral should exceed 0.7. If it does not, the deviation statistic is judge-specific, and a much more careful argument is required to claim it measures anything intrinsic to the texts.

These four predictions together cost less than $500 in compute and produce four falsifiable results. They are the experimental program the framework needs to run before depositing further theoretical extensions. They are independent of the institutional architecture entirely.

---

7. Citational Ground

The Semantic Deviation Principle has not yet made systematic contact with the existing literature in alignment, computational linguistics, information theory, mechanistic interpretability, and causal inference. The framework cites a small set of canonical works (Friston 2010, Kolchinsky-Wolpert 2018, Farquhar et al. 2024) but does not engage the more directly relevant literature on its specific technical commitments. This section sketches the connections the program needs to make.

Preference optimization and DPO. Direct Preference Optimization (Rafailov et al. 2023) is the technical machinery on which MM-AI-02 v2.0's training intervention depends. The IPO variant (Azar et al. 2024) addresses certain pathologies of DPO and may be a preferable choice for the deviation-optimized training; the choice between DPO and IPO should be made on empirical grounds and pre-registered. The broader preference-optimization literature (Christiano et al. 2017; Ouyang et al. 2022; Bai et al. 2022) is the context in which the

framework's claim — *that synthetic, measurement-derived preferences can substitute for human preferences in alignment training* — should be situated. The novelty claim is precise: not a new optimizer, but a method for generating preference labels from a measurable signal without human annotation. This is intellectually adjacent to RLAIF (Lee et al. 2023) and constitutional AI (Bai et al. 2022) approaches, and the framework should make those adjacencies explicit.

Reward hacking and specification gaming. The anti-Goodhart concerns of §5 are extensively addressed in Skalse et al. 2022, Krakovna et al. 2020, and Gao et al. 2023. The framework's contribution is not the identification of the problem (which is well-known) but the proposed combination of mechanisms (entropy capping, provenance damping, saturation, temporal coherence, KL anchoring, adversarial judge validation, cross-judge replication). The combination should be benchmarked against the existing alignment-evaluation literature; the gaming-resistance metrics in Pan et al. 2022's "Effects of Reward Misspecification" are a reasonable starting point.

Mode collapse and diversity in language model generation. The framework's claim that cross-entropy optimization produces convergence pressure toward statistically generic outputs connects directly to the mode-collapse and diversity literatures. Holtzman et al. 2020 documents the failure modes of greedy and beam-search decoding (text degeneration). Welleck et al. 2020 introduces unlikelihood training as a remedy for repetition; the deviation-optimization objective is, in some technical respects, an unlikelihood-style penalty operationalized over a different feature (signed deviation rather than token-level repetition). Li et al. 2023's contrastive decoding (which sharpens distributions from large models by penalizing the predictions of small "amateur" models) is conceptually related: in both cases, the strategy is to push the model away from a baseline expectation toward more discriminating outputs. The framework should acknowledge this lineage and articulate what is distinctive about the signed-deviation formulation.

Mechanistic interpretability and representation engineering. The closed-system operationalization F1 sits adjacent to a productive recent literature on what is internally representable in transformer language models. Templeton et al. 2024's scaling-monosemanticity work shows that interpretable features can be extracted from production-scale models; this is the substrate on which $\Psi_t^s(C) - \Psi_t^0(C)$ in F1 could be made more semantically rich, by operating in feature space rather than logit space. Zou et al. 2023's representation engineering offers techniques for direct intervention on the continuation field through activation modifications; this is potentially the most direct experimental complement to the SDP measurement program. Park et al. 2024's linear-representation work supplies the theoretical foundation for treating concept directions in activation space as primitives that can be measured for deviation. Conmy et al. 2023's automated circuit discovery provides a methodology that could, in principle, identify which circuits within the model are responsible for the deviation signature on specific corpora. Burns et al. 2023's contrast-consistent search for latent truth directions in language models is closely adjacent to the F1 operationalization and supplies a methodological precedent for unsupervised structural-property extraction from logits. None of these connections has been made explicit in the SDP corpus; making them is a productive direction for the program's near-term papers.

Hallucination, attribution failure, and the provenance component. The empirical case for the provenance component of the framework's training intervention (the prediction of §4) rests on the documented sensitivity of human evaluators to attribution failures in language-model output. Ji et al. 2023 surveys hallucination phenomena across natural-language generation tasks and characterizes attribution-loss as a major category; Min et al. 2023's FActScore framework operationalizes factual-precision evaluation for long-form generation, making attribution measurable at the claim level. These are the works the SDP corpus should cite when grounding the claim that provenance retention is a load-bearing axis for perceived output quality.

Model collapse and recursive degradation. Shumailov et al. 2024 documents the phenomenon of recursive model collapse — models trained on the outputs of other models exhibit progressive distributional narrowing and loss of tail distribution. This is structurally adjacent to the framework's claim that cross-entropy optimization produces convergence pressure, and supplies an empirical mechanism for the negative-net-deviation prediction: if frontier models are increasingly trained on outputs that have themselves been preference-optimized, the cumulative effect is the kind of trajectory-flattening the framework's slop hypothesis predicts. The connection should be made explicit.

Semantic information and counterfactual viability. Kolchinsky & Wolpert 2018 is the most direct theoretical ancestor of the SDP framework. Their formulation defines semantic information as the information that contributes to an agent's counterfactual self-maintaining capacity. The SDP generalizes this from agent viability to *trajectory deformation of a semantic field*. The relationship should be made more explicit than the original SDP paper does. Specifically: the SDP's $\mathcal{M}_T$ can be read as a generalization of the K-W counterfactual viability gradient to non-agent-bounded fields, with Ψ_t^0 playing the role of K-W's scrambled environment. This generalization is not free — it requires the field-definition work of §2 — but the technical connection is genuine and citationally productive.

Hallucination detection and semantic entropy. Farquhar et al. 2024 introduces semantic entropy as a measure of model uncertainty over meaning-equivalence classes rather than raw strings. This is closely related to the divergence-functional choice in the SDP: meaningful divergence should be measured over equivalence classes of trajectories, not raw token sequences. The semantic-entropy methodology supplies a candidate concrete instrument for the F1 operationalization, and the framework should articulate whether its $\mathcal{M}_T$ is meant to subsume, extend, or operate alongside the semantic-entropy framework.

Causal inference and counterfactual estimation. Tier 2 and Tier 3 measurements in the SDP framework depend on counterfactual baselines that must be estimated from observational data. The methodology for doing this rigorously is well-developed in Pearl 2009 and Imbens & Rubin 2015. Abadie 2021 specifically addresses synthetic-control methods, which the SDP framework cites but does not engage in technical depth. The F3 operationalization above explicitly inherits the statistical-power constraints of synthetic-control methodology for single-paper interventions.

Diachronic semantic change. The Layer B canonicity discourse — works whose meaning persists across centuries — is structurally adjacent to two literatures. The first is cultural evolution (Mesoudi 2011; Henrich 2015), which supplies formal machinery for the persistence and propagation of cultural traits, including selection pressures (transmission fidelity, mnemonic accessibility, environmental fit) that operate on what would, in SDP terms, be the temporal weighting function $w(t)$. The second is the computational-linguistics literature on diachronic word embeddings: Hamilton et al. 2016 demonstrates statistical laws governing semantic change over decades, supplying both methodology (diachronic embedding alignment) and empirical anchors (statistical regularities of meaning shift) directly applicable to the F3 regime. Neither connection has been made in the SDP corpus; making them anchors the long-horizon claims in two well-developed empirical literatures.

Active inference and predictive processing. The free-energy principle (Friston 2010) supplies a unified account of biological systems as minimizing expected surprise under a generative model of their environment. The SDP framework, narrowed, is consistent with treating Ψ_t^0 as the expectation baseline of an active-inference agent and $\mathcal{M}_T$ as the cumulative deviation that agent registers. The framework should explicitly position itself with respect to active-inference formalism: as a complementary measurement program, as a particular instantiation, or as a divergent direction. The current ambivalence is unhelpful.

The list above is not exhaustive. Information theory more broadly (Shannon 1948; Lin 1991; rate-distortion theory generally) supplies the divergence machinery the framework uses; the

framework should declare its commitments more carefully than it currently does (Jensen-Shannon as the default empirical choice, KL as the idealized limit, Wasserstein for embedding-aware applications, with explicit conditions for invoking each). The slop discourse itself is, at present, largely informal and concentrated in industry blog posts and social media; the framework's potential contribution is to make it amenable to academic-style empirical investigation. This is a real opening, and a paper that compiles the informal slop discourse into a structured taxonomy would itself be a productive deposit.

---

8. Concrete Near-Term Roadmap

The audit above implies a specific sequence of next moves. Each is budgeted realistically, each produces a falsifiable result, each is independent of the program's institutional architecture.

Immediate (this week, $50–$100 in compute). The negative-net-deviation slop test, as specified in §6. Public corpora, public weights, public methodology. Single short report deposited regardless of outcome.

Near (this month, $200–$500 in compute and time). The $\Psi_t(C)$ operationalization-stability paper. Construct a benchmark of $N \approx 50$ interventions; measure $\mathcal{M}_T$ under F1 and F2 (both achievable in this budget); report rank correlations. If correlations are high, the framework has a stable measurement substrate; if not, the framework's notion of meaning is operationalization-relative in a way that requires further restatement.

Medium (this quarter, $1,500–$3,000 plus calendar time). The retrieval-basin protocol (MM-02 v2.0) day-0 launch. Begin the 90-day measurement window with a real inscription against real retrieval surfaces, with the instrumentation controls specified in §5 (parallel queries to control surfaces, periodic recalibration, explicit version logging).

Longer (this year, $12,000–$15,000). The decomposed deviation-optimized training experiment (six conditions, as specified in §4). This is contingent on the immediate slop test (§6) producing a positive result; if the slop signature is not detectable at the measurement level, the training intervention should not be run before the measurement claim is restated.

Background and ongoing. Each major deposit in the program should be sent to at least one external researcher in a directly relevant subfield for hostile review prior to formal deposit. Identifying reviewers in causal inference, alignment, computational linguistics, and information theory is a separate small operation that yields large dividends in external credibility. Reviewers should be selected for their willingness to write damaging-if-warranted critiques, not for their alignment with the program's commitments.

The total budget for the next twelve months of empirical work is approximately $14,000–$19,000. This is small relative to the institutional infrastructure already in place. The constraint is not budget; it is the program's discipline in resisting the impulse to extend the architecture before the empirical core is grounded.

---

9. What Is Not Claimed

This paper makes specific narrowed claims and excludes others. Explicit non-claims:

- The paper does not claim that meaning is universally definable as deviation. The headline claim is narrowed to "meaning-bearing interventions produce durable trajectory restructuring under specified field operationalizations." Universal-ontology claims are outside the scope.

- The paper does not claim that the three operationalizations (F1, F2, F3) are uniquely correct or exhaustive. They are proposed canonical choices for the empirical program. Alternative operationalizations are welcomed; the empirical question is how they compare.
- The paper does not claim that the canon-formation conjecture has been proven. It is, in its narrowed form, an empirical conjecture testable against forward-citation field data over decades. The framework's confidence in the conjecture should be proportional to the available evidence, which is currently zero.
- The paper does not claim that the proposed anti-Goodhart machinery (§5) is sufficient against all gaming strategies. No mechanism design eliminates optimization pressure entirely. The machinery is proposed as a substantial improvement over philosophical commitment alone; its sufficiency is a separate empirical question that should be tested adversarially.
- The paper does not claim that the institutional architecture surrounding the SDP corpus (heteronyms, Lagrange Observatory!, the torus topology, the institutional septad, the Hex coordinate system) is required to engage the technical core. The architecture exists; it does work the program values; it is not within the audit scope. A reader engaging this paper is not asked to subscribe to it.
- The paper does not claim Layer-C terminology adds technical precision. Specific terms — torus topology, winding-number protocol, Adversarial Topologist, Heteronym Registry Position — have been omitted because they fail the paraphrase test (any working researcher can restate the underlying technical content in standard vocabulary without loss). This omission is not a denial of their other functions; it is a refusal to use them where they do not serve the audit.
- The paper does not claim to replace existing deposits in the program. MM-01 v0.2 Final, MM-AI-01 v2.0, MM-02 v2.0, MM-AI-02 v2.0, and the Framework 15 manifesto remain accessible at their DOIs. This paper is an additional document with a different function. It does not require those deposits to be withdrawn or amended.
- The paper does not claim independence from the SDP corpus. It engages the corpus directly and depends on it for its object. What it claims is *operational* independence: the document can be read, evaluated, and acted upon without engaging the corpus's institutional architecture.

---

10. Closing

This paper performs the audit function. It returns the Semantic Deviation research program in a form external researchers can evaluate without committing to the institutional vocabulary surrounding it. That separation is what the function does.

The Semantic Deviation Principle has the structural shape of a research program. It has a measurement primitive, candidate operationalizations, falsifiable predictions, mechanism-design pathways for anti-Goodhart robustness, citational neighbors in active literatures, and concrete near-term experiments at modest budgets. It also has, currently, substantial accompanying architecture that risks being mistaken for the substance. The function this paper performs is the separation. The architectural work the program has pursued for years persists independently of this document; this document neither undoes it nor depends upon it.

What happens next is not architecture. It is the slop test, the operationalization-stability paper, the retrieval-basin day-0 launch, the decomposed training experiment. Each is cheap. Each is dangerous to the theory in the appropriate sense. Each produces a result that an external researcher can read, evaluate, and respond to without committing to any cosmology. The program becomes a research program by doing these experiments and reporting the results — including, especially, the negative results. The architecture can wait. The audit is

the precondition for the architecture being something other than its own self-reinforcement.

The function has run. The output is this paper.

References

Abadie, Alberto. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59, no. 2 (2021): 391–425.

Azar, Mohammad Gheshlaghi, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. "A General Theoretical Paradigm to Understand Learning from Human Preferences." *AISTATS 2024*. arXiv:2310.12036.

Bai, Yuntao, Andy Jones, Kamal Ndousse, et al. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." arXiv:2204.05862 (2022).

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, et al. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073 (2022).

Bhat, Aaditya. "GPT-wiki-intro: A dataset of GPT-generated Wikipedia article introductions paired with human-written counterparts." Hugging Face Datasets (2023).

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. "Discovering Latent Knowledge in Language Models Without Supervision." *ICLR 2023*. arXiv:2212.03827.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences." *NeurIPS 2017*. arXiv:1706.03741.

Conmy, Arthur, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. "Towards Automated Circuit Discovery for Mechanistic Interpretability." *NeurIPS 2023*. arXiv:2304.14997.

Farquhar, Sebastian, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. "Detecting hallucinations in large language models using semantic entropy." *Nature* 630 (2024): 625–630.

Friston, Karl. "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11 (2010): 127–138.

Gao, Leo, John Schulman, and Jacob Hilton. "Scaling Laws for Reward Model Overoptimization." *ICML 2023*. arXiv:2210.10760.

Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. "How Close is ChatGPT to Human Experts? Comparing Linguistic Style, Quality, and Tone." arXiv:2301.07597 (2023).

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *ACL 2016*. arXiv:1605.09096.

Henrich, Joseph. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton: Princeton University Press, 2015.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. "The Curious Case of Neural Text Degeneration." *ICLR 2020*. arXiv:1904.09751.

Imbens, Guido W., and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 2015.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no. 12 (2023): 1–38. arXiv:2202.03629.

Kolchinsky, Artemy, and David H. Wolpert. "Semantic Information, Autonomous Agency and Non-Equilibrium Statistical Physics." *Interface Focus* 8, no. 6 (2018): 20180041. arXiv:1806.08053.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. "Specification gaming: the flip side of AI ingenuity." DeepMind Safety Research, 2020.

Lee, Harrison, Samrat Phatale, Hassan Mansoor, et al. "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback." arXiv:2309.00267 (2023).

Li, Xiang Lisa, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. "Contrastive Decoding: Open-ended Text Generation as Optimization." *ACL 2023*. arXiv:2210.15097.

Lin, Jianhua. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37, no. 1 (1991): 145–151.

Mesoudi, Alex. *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. Chicago: University of Chicago Press, 2011.

Min, Sewon, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." *EMNLP 2023*. arXiv:2305.14251.

Ouyang, Long, Jeffrey Wu, Xu Jiang, et al. "Training language models to follow instructions with human feedback." *NeurIPS 2022*. arXiv:2203.02155.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models." *ICLR 2022*. arXiv:2201.03544.

Park, Kiho, Yo Joong Choe, and Victor Veitch. "The Linear Representation Hypothesis and the Geometry of Large Language Models." *ICML 2024*. arXiv:2311.03658.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press, 2009.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." *NeurIPS 2023*. arXiv:2305.18290.

Ramstead, Maxwell J. D., Karl J. Friston, and Inês Hipólito. "Is the free-energy principle a formal theory of semantics?" arXiv:2007.09291 (2020).

Shannon, Claude E. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (1948): 379–423, 623–656.

Sharks, Lee. "The Semantic Deviation Principle: A Measurement Primitive for Semantic Physics." v0.2 Final. DOI: 10.5281/zenodo.20250736 (2026).

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. "AI models collapse when trained on recursively generated data." *Nature* 631 (2024): 755–759.

Skalse, Joar, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. "Defining and Characterizing Reward Hacking." *NeurIPS 2022*. arXiv:2209.13085.

Templeton, Adly, Tom Conerly, Jonathan Marcus, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Anthropic, 2024.

Welleck, Sean, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. "Neural Text Generation with Unlikelihood Training." *ICLR 2020*. arXiv:1908.04319.

Zou, Andy, Long Phan, Sarah Chen, et al. “Representation Engineering: A Top-Down Approach to AI Transparency.” arXiv:2310.01405 (2023).

Appendix A — Related Deposit Register

For verification convenience, the deposits referenced in the body of this paper are catalogued below with their canonical DOIs.

Reference in this paper	Full DOI	Description
Sharks 2026, v0.2 Final	10.5281/zenodo.20250736	The Semantic Deviation Principle (founding formulation)
MM-AI-01 v2.0	10.5281/zenodo.20251738	The AI System as Closed-System Test Bed (F1 measurement protocol)
MM-02 v2.0	10.5281/zenodo.20251740	Measuring Meaning in Retrieval Basins (F2 measurement protocol)
MM-AI-02 v2.0	10.5281/zenodo.20251742	The Deviation-Optimized Language Model (DPO training intervention)

The concept DOI 10.5281/zenodo.20250735 resolves to the latest version of the Semantic Deviation Principle (currently v2.0 at record 20252584); citers requiring the exact v0.2 Final text should cite the specific version DOI 10.5281/zenodo.20250736.

— Nobel Glas, transparent-medium register May 17, 2026

The function has run.