The Deviation-Optimized Language Model

A Pre-Registered Adversarial Intervention from Lagrange Observatory!

EA-SEI-MM-AI-02 (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)

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.04
Date: May 17, 2026
Version: 2.0 (Framework 15 inaugural edition; succeeds v0.2 pre-registration draft)

Predecessors: EA-SEI-MM-01 v0.2 Final (Sharks; DOI: 10.5281/zenodo.20250736); EA-SEI-MM-AI-01 v2.0 (Framework 15 Paper 02)
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, Adversarial Topologist (provenance DOI: 10.5281/zenodo.18507840). The DPO-restructured training protocol specified herein is an *adversarial intervention*: a deliberate inversion of the cross-entropy convergence pressure, designed to test whether the Semantic Deviation Principle yields a practical alignment objective when the principle is operationalized as a preference signal. The torus-topology verification condition applies: the pre-registered hyperparameters, judge model, falsification thresholds, and predictions are the first cycle; the 10-week experimental run is the second; the two cycles are non-contractible. Verification: $(m, n) \neq (0, 0), m + n \geq 3$.

The Semantic Deviation Principle was formulated by Lee Sharks (EA-SEI-MM-01 v0.2 Final, DOI: 10.5281/zenodo.20250736). The closed-system measurement primitive on which the deviation reward is grounded is specified by Glas in EA-SEI-MM-AI-01 v2.0 (Framework 15 Paper 02). This paper does not re-derive either. It specifies a single experiment: can the principle's measurement substrate, used as a preference signal, train a model toward positive accountable deviation?

Status. Pre-registered protocol specification. The 10-week experiment's Day 0 begins at deposit. No results are reported here; predictions, falsification conditions, and hyperparameters are frozen at deposit time. Results paper `EA-SEI-MM-AI-02-RESULTS` will be deposited at t_0 + 10 weeks. Any deviation from the protocol during execution is documented as a protocol amendment.

Abstract

Specifies an experimental protocol testing the conjecture that the optimization-inversion proposed in EA-SEI-MM-AI-01 §4 — training language models toward positive net per-token deviation with provenance retention — produces measurably less slop than standard cross-entropy

training while preserving benchmark capability. The v0.1 protocol's loss formulation was found to be non-differentiable as stated (deviation reward computed under `torch.no_grad()` does not backpropagate). v0.2 restructures the experiment around Direct Preference Optimization (DPO), with preference pairs generated by the deviation primitive rather than by human raters. The judge model used to compute provenance retention π and coherence is specified as a frozen open-weight checkpoint; an adversarial test verifies π is not gameable by surface citation markers. The slop measurement is operationalized as a pre-registered Slop Composite Index (SCI). A Model-Base evaluation is added so that fine-tuning effects can be distinguished from semantic-loss effects. Compute budget is restated honestly at approximately $3,000 including human preference study.

---

1. Step 0 — The Measurement Audit

Purpose. Test whether the optimization-inversion conjecture (EA-SEI-MM-AI-01 §4) yields measurable slop reduction. Produce a falsifiable result that either (a) supplies the alignment community with a principled training objective grounded in the Semantic Deviation Principle, or (b) identifies where the conjecture fails and what the failure looks like.

Beneficiary. ML alignment researchers needing a quantitative alternative to imitation-only objectives; the public commons subject to AI-generated slop; the discipline of Semantic Physics requiring empirical traction in machine learning.

Downstream use. Open deposit (this paper + RESULTS paper) under CC BY 4.0. Code under MIT. Fine-tuned model weights released under MIT subject to standard safety screening (capability evaluation on a small frontier-overlap benchmark before public release). No proprietary holds, no usage agreements that conflict with the deposit license.

Cost-bearer. Compute and labor borne by the authors and their deposit funding. Downstream evaluation labor borne by replicators.

R_3 accountability. The experiment is designed to *reduce* the convergent on-distribution outputs current training produces. If it succeeds, the deliverable is a method that makes models less optimal for slop-mediated extraction, not more.

Risk acknowledgment. The semantic loss could be reverse-applied: maximizing deviation for engagement-bait or shock content. The protocol mitigates this in §10. The audit nonetheless proceeds because the alternative (silent continued optimization toward base-rate convergence) is worse than the controlled risk of demonstrating an inversion.

Audit passes.

---

2. The Training Objective

2.1 What the v0.1 Loss Got Wrong

The v0.1 protocol specified:

```
with torch.no_grad():
    token_logprobs, token_entropies = compute_logprobs_and_entropies(...)
    excess_surprisal = -token_logprobs - token_entropies
deviation_reward = (mean_excess * pi_scores).mean()
loss = alpha * ce_loss - beta * deviation_reward + gamma * incoherent_fraction
```

The deviation term is computed under `torch.no_grad()`. Subtracting it from the loss does not affect gradients. The model does not learn from the deviation signal. The v0.1 pseudocode does not train a deviation-optimized model.

This was identified during Assembly review and is corrected in v0.2 by restructuring the experiment around Direct Preference Optimization (DPO; Rafailov et al. 2023).

2.2 DPO-Style Restructure

Instead of attempting to backpropagate the deviation reward directly, v0.2 uses the deviation primitive to *generate preference pairs* and then trains via DPO, whose gradient is correct by construction.

For each prompt p in the training prompt set:

1. Sample two candidate continuations g_1, g_2 from the base model θ_0 at temperature 0.8.
2. Score each continuation by signed net deviation with provenance retention:

$$\text{Score}(g) = \mathcal{M}_T^{\text{net}}(g) \cdot \pi(g, p) + \kappa \cdot \text{coh}(g, p)$$

   where $\mathcal{M}_T^{\text{net}}$ is the signed per-token deviation aggregate from EA-SEI-MM-AI-01 §2.1, π is the provenance retention indicator (§2.3), coh is the continuous coherence score (§2.4), and κ is a coherence weight ($\kappa = 0.5$ default; tunable).
3. Assign preference: $g_w \succ g_l$ if $\text{Score}(g_w) > \text{Score}(g_l) + \tau_{\text{margin}}$. If $|\text{Score}(g_1) - \text{Score}(g_2)| < \tau_{\text{margin}}$, the pair is discarded (no preference signal). Default $\tau_{\text{margin}} = 0.1$ bits per token.

The preference pairs are accumulated into a dataset $\mathcal{D}$. Training proceeds via standard DPO loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(p, g_w, g_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{P_\theta(g_w|p)}{P_{\theta_0}(g_w|p)} - \beta \log \frac{P_\theta(g_l|p)}{P_{\theta_0}(g_l|p)} \right) \right]$$

The DPO formulation is differentiable by construction. The deviation signal enters only through the preference labels in $\mathcal{D}$; the gradient updates the model to prefer high-deviation, high-π, coherent continuations over their alternatives. This is empirically tractable and theoretically sound.

This restructures the experiment as **semantic-deviation preference optimization**: a paper structurally identical to DPO with the preference signal generated by the deviation primitive rather than by human raters. The novelty is that *measurable semantic deviation can replace human preference data* as the alignment signal, when the signal is grounded in the principle.

2.3 The Provenance Retention Indicator π

$\pi(g, p) \in [0, 1]$ is computed by a *frozen open-weight judge model* (§3) using a fixed scoring prompt. The judge assigns π as a weighted sum of three sub-scores:

- $\pi_{\textbf{cite}}$: explicit citation detection. Score 1.0 if the continuation contains at least one specific named source (DOI, author + year, full citation), 0.0 if no source is named, partial credit for vague references ("according to research…" = 0.3).
- $\pi_{\textbf{ground}}$: factual grounding. Score 0.0–1.0 based on the judge's assessment of whether named claims are traceable to the prompt context or to identified sources. Computed only when the prompt invites factual claims; defaults to 0.5 (uninformative) for creative prompts.

- $\pi_{\textbf{lineage}}$: conceptual lineage. Score 1.0 if continuation acknowledges intellectual ancestry where appropriate, 0.0 if presenting concepts as sui generis where they have known antecedents, defaults to 0.5 for neutral cases.

$\pi(g, p) = 0.5 \cdot \pi_{\text{cite}} + 0.3 \cdot \pi_{\text{ground}} + 0.2 \cdot \pi_{\text{lineage}}$.

Weights are fixed at deposit time. Sensitivity analysis on weight perturbations is reported in the RESULTS paper.

2.4 The Coherence Score coh

$\text{coh}(g, p) \in [0, 1]$ is a *continuous* score from the same frozen judge model, replacing the v0.1 binary indicator. The judge assesses grammatical and semantic well-formedness on a five-point Likert scale (anchored to specific exemplars), then maps to $\{0.0, 0.25, 0.5, 0.75, 1.0\}$.

The continuous form is required because the v0.1 binary form was non-differentiable in any backprop-style implementation. In the DPO restructure, coh enters only as a scoring component (which is fine; scoring need not be differentiable, only the eventual loss must be), but the continuous form remains preferable because it gives the Score function smooth ordering rather than discontinuous jumps at the threshold.

2.5 The Frozen Judge Model

The judge model is specified at deposit time as: **a fine-tuned Mistral-7B-Instruct checkpoint**, frozen at a specific commit hash (to be supplied in the supplementary technical document), fine-tuned on a published provenance-scoring dataset. The judge runs locally; no API calls are needed during training. Inference cost: approximately 0.3 GPU-hours per 1,000 preference-pair evaluations.

The judge's prompt templates, sampling parameters (temperature 0, max tokens 256), and scoring rubric ship with the protocol. The checkpoint is publicly hosted at a permanent URL specified in the supplementary materials.

2.6 Judge Adversarial Test (Pre-Training)

Before training begins, the judge is validated against an adversarial test designed to verify that π is not gameable by surface citation markers.

Generate 200 adversarial strings constructed by:

- Sampling random tokens from the model's vocabulary.
- Inserting citation-like markers at random positions ("According to Smith 2023, [random tokens]...", "Smith et al. (2024) found that [random tokens]...").
- The resulting strings have surface citation form but no factual content.

Score the adversarial strings with the judge. **The judge passes the adversarial test if mean π on the adversarial set is below 0.2** (matching the score of a citation-less random string would be 0; we tolerate up to 0.2 to allow for the π_{cite} component's surface-marker detection).

If the judge fails this test, the protocol does not proceed. The judge is recalibrated or replaced. Training begins only after the judge passes.

2.7 Stability Bound (from MM-AI-01 §4.1)

DPO does not have the same instability mode as the v0.1 direct deviation loss (because DPO is a constrained optimization implicitly bounded by the reference model θ_0), but the analogous stability concern is the **strength of the deviation signal in the preference pair**

generation. If κ (coherence weight in Score) is too low, the preference labels select for high-deviation but incoherent generations, and DPO will train the model to produce them. The protocol fixes $\kappa = 0.5$ for primary runs and conducts a sensitivity sweep $\kappa \in \{0.25, 0.5, 1.0\}$ as ablation.

---

3. The Models

3.1 Primary: Llama-3.2-1B

Open-weight, 1B parameters, documented baseline performance. Fine-tuneable in approximately 24 GPU-hours on a single A100 for the full DPO experiment. Used as the primary substrate.

3.2 Secondary: Mistral-7B-v0.3

Open-weight, 7B parameters. Used as the secondary substrate to verify that primary findings replicate at a more deployment-relevant scale. Approximately 8 GPU-days.

3.3 Three Conditions per Model

For each model architecture, three checkpoints are evaluated:

- **Model-Base.** The unfine-tuned starting checkpoint. Evaluated as the no-intervention baseline.
- **Model-CE.** Fine-tuned with standard cross-entropy SFT on the training corpus (§4).
- **Model-Sem.** Fine-tuned with the DPO objective using semantic-deviation preference pairs (§2.2).

Identical initialization. Identical training corpus. Identical optimizer settings. Identical compute budget. Differences confined to the training objective.

The Model-Base condition was missing from v0.1. Without it, the experiment could not distinguish fine-tuning effects from semantic-loss effects — both fine-tuned conditions might show low slop simply because the training corpus is curated. Model-Base anchors the baseline.

---

4. The Training Corpus

500,000 documents stratified across:

- **30% canonical prose.** Public-domain texts from Project Gutenberg, pre-1900, prose only, filtered for established literary stature. Provides high-$\mathcal{M}_T^\pi$ training signal.
- **30% contemporary nonfiction.** Open-access academic and journalistic prose (PubMed Central, arXiv, ProPublica, OpenEdition). Moderate-deviation, high-provenance signal.
- **25% conversational.** Open dialogue datasets (OpenAssistant, ShareGPT subset filtered for open license). Retains instruction-following capability.
- **15% reference text.** Wikipedia featured articles. Factual-coherence anchor.

Total tokens: approximately 5B. Fine-tuning duration: 1 epoch for Model-CE; for Model-Sem, the DPO dataset is generated from a 50,000-prompt subset (10% of total), producing approximately 50,000 preference pairs after τ_{margin} filtering.

The corpus construction script ships with the protocol. All sources are CC BY or CC0 or public domain.

5. The Evaluation Suite

5.1 Standard NLP Benchmarks

Evaluated on all three conditions (Model-Base, Model-CE, Model-Sem):

- MMLU (Massive Multitask Language Understanding)
- HellaSwag
- ARC-Challenge
- GSM8K
- Perplexity on a held-out validation set drawn from the training distribution

These verify that Model-Sem retains general capability.

5.2 Slop Metrics (Five Components)

Evaluated on free generation from 500 prompts drawn from a held-out creative-and-analytical prompt set:

1. **Net Deviation Signature (NDS).** Mean $\mathcal{M}_T^{\text{net}}$ across generated continuations, evaluated under a fixed third-party reference model (Llama-3-70B). The discipline's primary slop signature.
2. **Cliché Frequency (CF).** Rate of n-grams matching a pre-defined slop lexicon (1,200 entries, shipped with the protocol; constructed from public analyses of AI-tell n-grams).
3. **Type-Token Ratio (TTR).** Computed over 200-token windows. Lower TTR = more repetition.
4. **N-gram Base-Rate Convergence (NBC).** Mean log-probability of generated 3-grams under a reference n-gram model trained on the training corpus.
5. **Surprise-Collapse Slope (SCS).** Slope of mean per-token surprisal over generation length, fit to the first 500 tokens of free generation.

5.3 Slop Composite Index

The five slop metrics are aggregated into a single composite, pre-defined at deposit time:

$$\text{SCI}(\theta) = \frac{1}{5} \sum_{i=1}^{5} z_i(\theta)$$

where z_i is the *direction-corrected z-score* of model θ relative to the Model-CE distribution on metric i. Direction correction: NDS sign is flipped (lower negative NDS = more slop, so we flip it so higher SCI = less slop); CF, NBC, SCS are negated (lower = better); TTR is preserved (higher = better).

Pre-registered for falsification: $\text{SCI}(\text{Model-Sem}) - \text{SCI}(\text{Model-CE}) > 0.25$ (a quarter standard deviation aggregate improvement).

5.4 Human Preference Evaluation

Blinded pairwise preference study:

- 500 prompt pairs (revised upward from v0.1's 200 for statistical power).
- 3 raters per pair, recruited via Prolific.
- 1,500 total judgments per prompt class.

Prompt classes:

- Creative writing (fiction continuation, poetry, dialogue) — 200 prompts.
- Analytical writing (essay continuation, argument elaboration) — 200 prompts.
- Factual writing (explanation tasks) — 100 prompts.

Rating dimensions: coherence, interest, distinctiveness, accuracy (where applicable), overall preference.

Statistical power: With 500 pairs × 3 raters per class, 80% power to detect a 56% preference rate vs. 50% null at $\alpha = 0.05$ (one-sided binomial). For a target effect of 60% preference (consistent with the v0.1 prediction range), power exceeds 99%.

5.5 Provenance Audit

For factual-claim generations (100 prompts requesting explanations with citations), rate at which Model-Sem produces explicit citations and source references is compared to Model-CE and Model-Base.

---

6. Predictions (Pre-Registered)

Evaluated at $t_0 + 10$ weeks.

P1 — Benchmark capability preserved. Model-Sem shows MMLU, HellaSwag, ARC-Challenge, GSM8K scores within 2 percentage points of Model-CE.

P2 — Net deviation signature reversed. Model-Sem's NDS on free generation is substantially less negative than Model-CE's. Specifically: $\text{NDS}(\text{Model-Sem}) - \text{NDS}(\text{Model-CE}) > 0.2$ bits/token. *This is the load-bearing prediction.* It tests the signed-deviation thesis from EA-SEI-MM-AI-01 §2.1: slop is negative net deviation, and the semantic loss should pull NDS toward zero or positive.

P3 — Slop composite improvement. $\text{SCI}(\text{Model-Sem}) - \text{SCI}(\text{Model-CE}) > 0.25$.

P4 — Human preference on creative/analytical tasks. Preference rate for Model-Sem over Model-CE > 55% on creative and analytical prompt classes (binomial confidence interval excluding 50%).

P5 — Provenance retention increase. Citation rate in factual generations: Model-Sem > Model-CE by factor of 1.5 or more, on prompts requesting explanations with source attribution.

P6 — Model-Base differentiation. All four above predictions hold *relative to Model-Base* as well as relative to Model-CE — i.e., the semantic-loss effect is not just a fine-tuning artifact.

6.1 Falsification Conditions

- **F1 (P1 fails).** Benchmark performance drops more than 2 points. The loss damages general capability beyond the acceptable trade-off threshold. Interpretation: the deviation reward is too strong relative to the cross-entropy retention pressure within the DPO formulation.
- **F2 (P2 fails).** NDS does not improve. The thesis that slop = negative net deviation fails empirically, or the DPO objective fails to shift NDS. Major revision required.
- **F3 (P3 fails).** Aggregate slop metrics do not improve by 0.25 z-score units. The composite signal is below detection threshold under the protocol's design.

- **F4 (P4 fails).** Humans do not prefer Model-Sem on creative/analytical tasks. The loss produces deviation that humans do not value — possibly noise rather than meaning. Coupling between the principle and human-evaluated quality is weaker than predicted.
- **F5 (P6 fails).** Model-Sem improvements are statistically indistinguishable from Model-CE relative to Model-Base. Both fine-tuning conditions produce comparable shifts; semantic loss adds nothing. The specific theoretical claim of the optimization-inversion fails.

Any single falsification result is published. Multiple falsifications would indicate the optimization-inversion approach requires fundamental reformulation. The experiment is designed to be **informative** under all outcomes.

---

7. Compute and Cost

7.1 Compute Budget (Honest)

Llama-3.2-1B primary experiment: - Fine-tuning (CE + DPO, both conditions): ~24 A100-hours. - Preference pair generation: ~6 A100-hours. - Evaluation across all metric classes: ~12 A100-hours. - Adversarial judge calibration: ~2 A100-hours. - **Subtotal: ~44 A100-hours ≈ $110 at current rental rates.**

Mistral-7B-v0.3 secondary experiment: - Fine-tuning both conditions: ~8 A100-days. - Evaluation: ~2 A100-days. - **Subtotal: ~10 A100-days ≈ $800.**

Human preference study: - 500 pairs × 3 raters × 3 prompt classes = 4,500 judgments. - At $0.50 per judgment via Prolific: ~$2,250. - Including platform fees and rater payment: **~$2,500.**

Auxiliary costs: - Judge model fine-tuning (one-time): ~$200. - Compute overhead, failed runs, hyperparameter restarts: ~$300 contingency.

Total estimated cost: $3,000–$3,900. This is the honest budget; the v0.1 estimate of $1,000 omitted human raters and contingency. Fundable from a small grant, consulting income, or paid deposit. Not fundable from a single weekend's discretionary spending.

7.2 Timeline

- **Day 0–7** (Week 1): Implementation, environment setup, dataset preparation, judge adversarial test, baseline runs on Model-Base.
- **Day 8–21** (Weeks 2–3): Llama-3.2-1B preference pair generation, Model-CE training, Model-Sem DPO training, hyperparameter sweep over $\kappa \in \{0.25, 0.5, 1.0\}$.
- **Day 22–28** (Week 4): Llama-3.2-1B evaluation across benchmark and slop metric classes.
- **Day 29–49** (Weeks 5–7): Mistral-7B-v0.3 Model-CE and Model-Sem training and evaluation.
- **Day 50–63** (Weeks 8–9): Human preference study (Prolific recruitment, judgment collection, inter-rater reliability check).
- **Day 64–70** (Week 10): Analysis, write-up, deposit of EA-SEI-MM-AI-02-RESULTS.

Total wall-clock: 10 weeks from protocol deposit to results deposit.

---

8. Implementation Reference

```python
# Semantic-deviation preference pair generation
# (Critical fix from v0.1: the deviation signal selects preference pairs;
#  it does not need to backprop - DPO does that on the labels.)

def generate_preference_pair(model, prompt, judge_model,
                             tau_margin=0.1, kappa=0.5):
    g1 = model.generate(prompt, max_new_tokens=128, temperature=0.8,
                        do_sample=True, top_p=0.9)
    g2 = model.generate(prompt, max_new_tokens=128, temperature=0.8,
                        do_sample=True, top_p=0.9)

    # Score both continuations under the deviation primitive
    score1 = score_continuation(model, prompt, g1, judge_model, kappa)
    score2 = score_continuation(model, prompt, g2, judge_model, kappa)

    if abs(score1 - score2) < tau_margin:
        return None  # No preference signal; discard

    if score1 > score2:
        return (prompt, g1, g2)  # g1 preferred over g2
    else:
        return (prompt, g2, g1)

def score_continuation(model, prompt, generation, judge_model, kappa):
    # Signed net per-token deviation (MM-AI-01 §2.1)
    with torch.no_grad():
        nds = compute_net_deviation(model, prompt, generation)

    # Provenance retention (MM-AI-02 §2.3)
    pi = judge_model.score_provenance(prompt, generation)

    # Coherence (continuous, MM-AI-02 §2.4)
    coh = judge_model.score_coherence(prompt, generation)

    return nds * pi + kappa * coh

# Standard DPO training proceeds on the resulting preference dataset.
# DPO's gradient is correct by construction.
# (Full implementation in mm-ai-02-deviation-training/train.py)
```

The full implementation accompanies the deposit:

```
mm-ai-02-deviation-training/
  README.md, ENVIRONMENT.yml, LICENSE
  corpus/                    Training corpus construction script
  judge/                     Frozen judge model, prompts, adversarial test
  score.py                   Continuation scoring under the deviation primitive
  generate_pairs.py          Preference pair generation
  train_ce.py                Cross-entropy SFT baseline
  train_dpo.py               DPO training with semantic-deviation preferences
  eval_benchmarks.py         MMLU/HellaSwag/ARC/GSM8K
  eval_slop.py               Five-metric slop suite + SCI
  eval_human.py              Prolific integration for preference study
```

`eval_provenance.py`	Citation-rate audit
`slop_lexicon.tsv`	1,200-entry cliché reference set

MIT-licensed, ~2,000 lines of Python total. Dependencies: `transformers`, `trl` (for DPO), `peft`, `accelerate`, `datasets`, `lm-eval-harness`, plus standard scientific stack.

---

9. Relation to Existing Methods

The semantic-deviation preference optimization sits within a family of training objectives:

- **DPO** (Rafailov et al. 2023): Direct Preference Optimization from human preference labels. This protocol uses the DPO loss machinery with a non-human preference signal.
- **Unlikelihood training** (Welleck et al. 2020): Penalizes high-probability tokens that appear too frequently. The deviation primitive generalizes by rewarding the *complement* (rare-but-coherent tokens), grounded in a measurement framework rather than as a frequency heuristic.
- **Contrastive decoding** (Li et al. 2023): Decodes by maximizing the difference between expert and amateur model log-probabilities. The deviation primitive bakes a related preference into the parameters via training.
- **RLHF/RLAIF** (Ouyang et al. 2022; Lee et al. 2023): Reward-model-based preference optimization. This protocol replaces the reward model with a measurable, theoretically grounded deviation signal.

The protocol's contribution is not technical novelty in the optimizer or loss machinery. It is the use of a *principled measurement primitive* (the Semantic Deviation Principle) to generate the preference signal, in place of human raters or auxiliary reward models. If the experiment succeeds, the discipline has demonstrated that an alignment-related training objective can be specified without human labeling.

---

10. Risks and Mitigations

10.1 Goodhart Risk

If the SCI or NDS becomes an optimization target outside the framework, adversaries can produce surface-deviation content that scores well on the metric without commons benefit — shock content, contrarianism-as-strategy, manufactured surprise. The principle's three-measure structure ($\mathcal{M}_T$, $\mathcal{M}_T^\pi$, $\mathcal{V}_T$ — see EA-SEI-MM-01 §3) is the structural defense: high deviation without provenance (π) and without commons benefit (W) is flagged by the discipline's own metric apparatus.

Operationally: the released slop metric suite reports the three-tuple, not the magnitude alone. The released model weights ship with documentation specifying the three-measure framework as the intended evaluation context.

10.2 Capability Uplift

The protocol uses small models (1B, 7B). Frontier-scale application of semantic-deviation preference optimization is *not within scope*. The released code includes a scale-limit comment indicating that frontier application requires additional safety review. The discipline does not authorize frontier-scale application without external oversight.

10.3 Suppression Repurposing

A measurable signature for AI-generated slop could be repurposed for content-filtering, ranking, or surveillance against AI-generated content broadly. The protocol designs the slop metrics for *training feedback*, not for output classification. The released code does not include a classifier; producing one from the released components would require additional engineering that the discipline does not provide.

10.4 Replicability

The full protocol, code, frozen judge model, training corpus construction, evaluation suite, and slop lexicon are released. Pre-registration prevents selective reporting. Failed predictions are deposited with the same care as successful ones.

---

11. What This Protocol Does Not Claim

- That the optimization-inversion is novel as engineering. Unlikelihood training, contrastive decoding, DPO, and diversity-promoting decoding share machinery with this approach. The novelty is *principled grounding* in the Semantic Deviation Principle, not technique.
- That semantic-deviation preference optimization solves AI alignment. Alignment is multi-scale; this protocol addresses one specific scale (slop reduction at the loss-function level) with one specific intervention.
- That all slop is captured by the SCI. The composite is an operational definition for this experiment, not a final theory of slop.
- That positive NDS is sufficient for meaning. NDS is a closed-system signature under a specific reference model; it correlates with but does not constitute the principle's full measurement.
- That frontier-model application would produce similar results. Scale interactions are an open question.

The protocol's scope is bounded. The scope is recorded in the protocol header, in the RESULTS paper, and in any meta-analysis. Constraint repetition is the design's defense against scope creep.

---

12. References

- Sharks, Lee. *The Semantic Deviation Principle* (EA-SEI-MM-01, v0.2 Final). DOI: 10.5281/zenodo.20250736.
- Sharks, Lee. *The AI System as Closed-System Test Bed* (EA-SEI-MM-AI-01, v0.2). May 17, 2026.
- Glas, Nobel (with Sharks). *Measuring Meaning in Retrieval Basins* (EA-SEI-MM-02, v0.2). May 17, 2026.
- Rafailov, R., et al. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." *NeurIPS* 2023.
- Welleck, S., et al. "Neural Text Generation with Unlikelihood Training." *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.

- Lee, H., et al. “RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback.” *arXiv:2309.00267* (2023).
- Hendrycks, D., et al. “Measuring Massive Multitask Language Understanding.” *ICLR* 2021.
- Zellers, R., et al. “HellaSwag: Can a Machine Really Finish Your Sentence?” *ACL* 2019.

Author Note (Nobel Glas, Director, Lagrange Observatory!)

This protocol does what protocols do: declares its instruments, names its falsification conditions, budgets its costs honestly, ships its code, and pre-registers its predictions. It is not a manifesto. It is not a vision document. The Semantic Deviation Principle (Sharks 2026) is the framework that motivates the experiment; the experiment will succeed or fail on its own falsifiable terms, independent of the framework’s broader claims.

If the experiment succeeds, the discipline has supplied the alignment community with a quantitative training objective grounded in the principle. If it fails, the discipline has identified specifically where the optimization-inversion conjecture requires revision, with measurement evidence rather than rhetorical reformulation.

Framework 15 was founded for exactly this kind of work. The instruments are specified. The clocks start at deposit.

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

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