Measuring Semantic Deviation: Operationalizations, Experiments, and Falsification Conditions for a Theory of Meaning as Field Deformation

Nobel Glas

2026-05-17

Measuring Semantic Deviation: Operationalizations, Experiments, and Falsification Conditions for a Theory of Meaning as Field Deformation

Nobel Glas[1] ORCID: 0009-0000-1599-0703

[1] Nobel Glas is a heteronym of Lee Sharks, adopted for this measurement program to signal that the empirical protocols are designed for independent replication regardless of the reader's engagement with the broader Crimson Hexagonal Archive. Correspondence and ORCID are maintained through Lee Sharks.

Abstract

This paper presents a measurement program for the Semantic Deviation Principle, which defines meaning as the time-integrated divergence a sign induces from the most probable trajectory of a semantic field. The principle yields a scalar quantity — raw semantic magnitude — computable wherever the field admits a probability distribution and a divergence functional, extending the Bar-Hillel and Carnap (1953) program for semantic information into distributional and temporal domains. We specify two executable operationalizations of this quantity: (F1) closed-system trajectory deviation within a frozen language model, where the counterfactual baseline is read directly from logits, building on surprisal theory (Hale 2001; Levy 2008) while decomposing it into signed deviation from conditional entropy; and (F2) retrieval response deviation across external AI search surfaces over a 90-day prospective window. A third operationalization (F3, citation graph deviation) is described as a long-horizon complement. We identify signed per-token deviation as a tractable proxy for F1 and derive a falsifiable prediction: AI-generated text exhibits statistically significant negative mean signed deviation relative to matched human text — a claim testable with existing corpora and complementary to log-probability curvature methods (Mitchell

et al. 2023). We describe a Direct Preference Optimization (Rafailov et al. 2023) experiment that uses the deviation primitive to generate preference pairs, extending the RLHF lineage (Christiano et al. 2017; Ouyang et al. 2022) by replacing human preference data with a measurable semantic signal. We specify six mechanism-design protections against Goodhart collapse (Manheim & Garrabrant 2019), each with operational calibrations. We pre-register the cheapest dangerous test with named datasets, frozen reference checkpoints, and statistical procedures. Total budgeted program: approximately $14,000–$19,000 across twelve months. Results deposited regardless of outcome.

---

1. Introduction

The question of whether meaning admits measurement has been posed formally since Bar-Hillel and Carnap's 1953 theory of semantic information, which defined semantic content as the set of possible states a proposition excludes. Kolchinsky and Wolpert (2018) linked semantic information to dynamical consequences, defining it as mutual information between an agent's internal state and its environment that is causally relevant to viability — the first formalization tying semantic content to trajectory effects rather than static probability assignments. The present framework shares with Kolchinsky and Wolpert the intuition that semantic content is measured by dynamical consequences, but replaces their viability-conditioned mutual information with a field-level divergence integral, removing the agent-boundedness requirement and enabling measurement in non-agentive semantic fields (publication corpora, retrieval surfaces, language model continuations).

The Semantic Deviation Principle (Sharks 2026, DOI: 10.5281/zenodo.20250736) proposes:

> *Meaning is the temporal deviation a sign induces from the field's probable evolution.*

This is not Shannon surprisal, which measures the unlikelihood of a sign at the moment of its appearance: $I(s) = -\log P(s)$. A random string has high Shannon surprisal but near-zero semantic magnitude, because it produces no durable deformation of any field's trajectory. The distinction — between instantaneous improbability and sustained trajectory restructuring — is the principle's load-bearing claim.

The program described here asks what happens when you try to compute the integral. The answer, across two primary operationalizations, is that the computation is tractable in several regimes, the resulting quantities are experimentally discriminable, and the framework produces falsifiable predictions testable with modest resources.

---

2. The Semantic Deviation Principle

2.1 Raw Semantic Magnitude

Let C denote a semantic context or field, s a sign introduced at time t_0, $\Psi_t^0(C)$ the probability distribution over future semantic states without s, $\Psi_t^s(C)$ the distribution with s, D a divergence functional, and $w(t)$ a temporal weighting function. The raw semantic magnitude of s over horizon T is:

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

When $w(t)$ is normalized ($\int w(t)\, dt = 1$), the magnitude retains the units of D: bits under Jensen-Shannon with $\log_2$ — a mean deformation intensity. When unnormalized, it is accumulated divergence-over-time: total semantic work in bit-days or bit-years. Both forms are legitimate and must be declared with each measurement.

Jensen-Shannon divergence is the default for empirical work: symmetric, bounded in $[0, \log 2]$ bits, finite when supports differ. KL divergence is the asymmetric, idealized limit. Wasserstein distance respects state-space geometry and is preferred when state distance carries semantic content. The geometric form ($\mathcal{M}_T = \int w(t) \|\Psi_t^s - \Psi_t^0\|^2\, dt$; cf. information geometry, Amari 2016) is noted for completeness; all empirical work in this paper uses the distributional form.

2.2 Provenance-Resolved Magnitude

Let PER denote the Provenance Erasure Rate — the fraction of provenance-bearing relations (authorial lineage, conceptual ancestry, source attribution) severed during transmission, connecting to the broader literature on attribution in NLG (Bohnet et al. 2022; Rashkin et al. 2023) and factual precision (Min et al. 2023). Provenance resolution modulates the *magnitude* of the deviation, not its sign:

$$\mathcal{M}_T^{\pi}(s \mid C) = |\mathcal{M}_T(s \mid C)| \cdot (1 - \text{PER}) \cdot \text{sgn}(\mathcal{M}_T)$$

When PER = 0, the deformation is fully accountable. When PER = 1, the deformation persists but its origin is unrecoverable — orphan deformation, present and effective and unattributable. For later use in the broader semantic-economy framework, we denote the accountable share of raw magnitude by $\oint = 1 - \text{PER}$.

A third derived measure — normative semantic value $\mathcal{V}_T = \mathcal{M}_T^{\pi} \cdot W$ — estimates whether accountable deformation enriches the commons or extracts from it. W is a sketch, not yet an instrument; this paper works exclusively with $\mathcal{M}_T$ and $\mathcal{M}_T^{\pi}$.

2.3 The Counterfactual Baseline

The principle requires Ψ_t^0. We adopt the tiered approach standard in causal inference (Pearl 2009; Imbens & Rubin 2015):

Tier 1 (tractable): Prospective intervention studies. Pre-register query set, divergence functional, and horizon. Record baseline. Introduce s. Observe and integrate.

Tier 2 (difficult): Natural experiments with synthetic controls (Abadie 2021). Identify comparable fields, one exposed to s, one not. Report with uncertainty bounds.

Tier 3 (approximable): Historical cases. Upper-bound by maximum-entropy Ψ_t^0; lower-bound by nearest-neighbor trajectory. The diachronic word embedding methods of Hamilton, Leskovec, and Jurafsky (2016) are relevant as empirical estimates of trajectory change in historical semantic fields.

The experimental program in this paper operates entirely at Tier 1.

---

3. Two Primary Operationalizations

3.1 F1 — Closed-System Continuation Field

Field: The conditional next-token distribution of a fixed language model checkpoint θ.

The counterfactual advantage. A trained language model at inference time is observationally closed: no new data enters, no weights update. The baseline Ψ_t^0 is the model's own conditional distribution, read directly from logits. The model's conditional distribution is a calibrated proxy for the external semantic field; F1 measures deviation from this proxy, not from the field itself. Convergent measurements across multiple reference models (P4, §7) strengthen inference to the field.

Per-token deviation (tractable proxy). For a sequence $x_{1:T}$ evaluated against frozen θ:

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

The first term is standard token surprisal — the quantity studied in psycholinguistic models of processing difficulty (Hale 2001; Levy 2008; Smith & Levy 2013). The second is the conditional entropy, which Meister, Cotterell, and Vieira (2021) use for the uniform information density hypothesis. The difference is signed: positive δ_t indicates a deviation event (the token is more surprising than the model's baseline expectation); negative δ_t a convergence event (more probable than expected).

This signed decomposition is related to but distinct from the log-probability curvature used in DetectGPT (Mitchell et al. 2023), which asks whether text sits at a local maximum of the model's log-probability surface. Signed deviation asks whether each token deviates from or converges toward the model's conditional entropy — a different geometric property of the probability landscape.

The mean signed per-token deviation is denoted $\bar{\delta}$. Throughout this paper, $\bar{\delta} = \mathcal{M}_T^{\text{net}} = \frac{1}{T}\sum_{t=1}^{T} \delta_t$. The absolute aggregate $\mathcal{M}_T^{\text{abs}} = \frac{1}{T}\sum_{t=1}^{T} |\delta_t|$ is reported as a secondary robustness check. Units: bits per token. $\bar{\delta}$ is the operative primitive for F1.

Signed per-token deviation is not identical to raw semantic magnitude. It is the closed-system local proxy tested in this paper: a token-level observable derived from the same deviation logic, while the trajectory-distribution form below remains the direct analog of the general principle. The two measures are expected to correlate strongly when an intervention produces consistent signed deviation across positions; they diverge when deviations oscillate in sign. The cheapest dangerous test (§7) uses the per-token form for computational tractability; the full trajectory form is reserved for validation studies.

Closed-system trajectory deviation (load-bearing form). The direct analog of $\mathcal{M}_T$:

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

Estimation proceeds via sampled rollout feature distributions, connecting to the distributional approach used in MAUVE (Pillutla et al. 2021) — though MAUVE measures distributional similarity between corpora while trajectory deviation measures shift induced by a specific intervention.

Provenance-resolved variant: The provenance retention indicator π modulates the magnitude of the per-token deviation: $\delta_t^\pi = |\delta_t| \cdot \pi_t \cdot \text{sgn}(\delta_t)$, where π_t is evaluated on the sequence including any provenance markers. High δ_t^π requires both positive signed deviation and intact provenance.

Parameter	Commitment
Divergence	KL over softmax logits (per-token); JS (trajectory)
Weighting	Uniform over positions (normalized)
Horizon	512 tokens default
Baseline	Read from logits (model as calibrated proxy)

3.2 F2 — Retrieval Response Field

Field: Response distributions of AI retrieval surfaces to a fixed query set, sampled over a 90-day window.

Surface taxonomy. Retrieval-mediated surfaces (Class R: Google AI Overview, Perplexity, ChatGPT with browsing) are separated from parametric surfaces (Class P: Claude, Gemini, ChatGPT without browsing). The headline metric $\mathcal{M}_T^{\text{retrieval}}$ uses Class R only; Class P is reported separately. Pooling confounds retrieval-basin deformation with training-data drift.

Measurement. Responses are captured through surface-appropriate collection methods: official APIs where available, and pre-registered browser- or SERP-level capture protocols where the target surface exposes no research API. A frozen extractor model (open-weight, documented commit hash) extracts named entities (spaCy with Wikidata QID resolution), claims (SPO triples), and citations (URLs, DOIs, named references). Divergence:

$$D_q(t_i) = D_{JS}\Big(R^s_{t_i}(q) \Big\| R^0_{t_0}(q)\Big)$$

with Laplace smoothing ($\alpha = 1$). Robustness: divergence under three representations (raw, embedding-smoothed, human-audited subsample); Pearson $r > 0.7$ required (Deutsch, Doshi, & Roth 2022).

Three-condition control. S (full identity), S^* (blank identity — ORCID omitted, author listed as "Anonymous"), S^{**} (plausible synthetic identity: single-purpose ORCID, realistic fabricated name, no prior deposits, no institutional affiliation, designed to be indistinguishable from an early-career researcher's first deposit). This separates content effects from identity-scaffolding effects.

Parameter	Commitment
Divergence	JS with Laplace smoothing ($\alpha = 1$)
Horizon	90 days; measurements at t_0, 7d, 28d, 84d
Baseline	Pre-intervention capture

3.3 Future Operationalization: Citation Graph Fields (F3)

Forward-citation distributions over a paper corpus (OpenAlex, Semantic Scholar) provide a long-horizon complement. Divergence: JS over topic-cluster distributions with regularized inverse-time weighting $w(t) = 1/(1 + t - t_0)$.

Statistical-power constraints are severe: single-paper interventions are typically underpowered within a 12-month window (Waltman 2016; Hicks et al. 2015). F3 is viable for aggregate interventions or with Bayesian hierarchical pooling, and is deferred to a follow-up study. No F3 predictions are pre-registered in this paper.

---

4. Machine-Output Convergence as Negative Deviation

4.1 The Cross-Entropy Argument

Standard language model training minimizes cross-entropy: $\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{T}\sum_t \log P_\theta(x_t \mid x_{<t})$. This drives the model toward the training corpus's base-rate continuations — the phenomenon Holtzman et al. (2020) documented as neural text degeneration and addressed with nucleus sampling (a generation-time intervention), and that Welleck et al. (2020) addressed with unlikelihood training (a training-time intervention).

Under the principle, this base-rate convergence has a specific numerical signature: $\bar{\delta} < 0$ — text that actively pulls toward the model's base rate, each token more probable than the conditional entropy expects. A second regime exists: temperature slop, where high-temperature sampling produces high $\mathcal{M}^{\text{abs}}_T$ without provenance — text that surprises but is unmoored from any source, connecting to the hallucination taxonomy of Ji et al. (2023) and the attribution failure modes of Min et al. (2023).

4.2 The Falsifiable Claim

> AI-generated text exhibits statistically significant negative mean signed deviation $\bar{\delta}$ relative to matched human-written text, computed against a frozen open-weight reference model.

This is distinct from existing detection methods. DetectGPT (Mitchell et al. 2023) uses log-probability curvature under random perturbation. Watermarking (Kirchenbauer et al. 2023) embeds statistical signatures during generation. Classifier-based methods face adversarial evasion (Sadasivan et al. 2023). Our approach measures a distributional property of the text against a reference model's conditional entropy — no perturbation, no watermark, no trained classifier. This makes it complementary and potentially more robust to adversarial evasion.

The connection to model collapse is direct: Shumailov et al. (2024) showed that training on recursively generated data produces progressive distributional collapse; Alemohammad et al. (2023) formalize this as self-consuming generative models. Under the deviation framework, model collapse is the progressive convergence of Ψ_t^s toward Ψ_t^0 — the field losing its capacity for deviation — and $\bar{\delta}$ provides a scalar measure of the severity.

---

5. The Training Intervention

The training intervention is not required to validate the measurement principle. It is included because a valid deviation primitive should be usable not only diagnostically but operationally: it should generate a preference signal whose downstream effects can be tested.

The RLHF lineage — from Christiano et al. (2017) through Ziegler et al. (2019), Stiennon et al. (2020), Ouyang et al. (2022, InstructGPT) — demonstrates that human preference signals can steer language model behavior. DPO (Rafailov et al. 2023) achieves this without an explicit reward model by optimizing directly against preference labels; DPO was chosen over PPO (Schulman et al. 2017) for simplicity, computational efficiency, and the absence of a separate reward model. IPO (Azar et al. 2024) and KTO (Ethayarajh et al. 2024) offer further simplifications.

Our experiment extends this lineage by asking: can the deviation primitive replace human preference data as the alignment signal?

5.1 Preference Pair Generation

For each prompt p, sample two continuations g_1, g_2 from base model θ_0 at temperature 0.8. Score each by:

$$\text{Score}(g) = \bar{\delta}(g) \cdot \pi(g, p) + \kappa \cdot \text{coh}(g, p)$$

The provenance retention indicator $\pi \in [0, 1]$ is a weighted sum of citation detection (π_{cite}, 0.5), factual grounding (π_{ground}, 0.3), and conceptual lineage (π_{lineage}, 0.2),

scored by a frozen judge model (Mistral-7B-Instruct, documented commit hash). These weights are pre-registered defaults; a sensitivity analysis varying each by ±50% is planned for the decomposed follow-up. The coherence score coh $\in [0, 1]$ is a continuous five-point Likert mapping from the same judge. Default $\kappa = 0.5$.

Preference: $g_w \succ g_l$ if Score difference exceeds $\tau_{\text{margin}} = 0.1$ bits/token. Pairs below margin discarded.

5.2 DPO Training

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

The deviation signal enters through preference labels; the gradient is correct by construction.

5.3 Three Conditions

For each architecture (primary: Llama-3.2-1B; secondary: Mistral-7B-v0.3): **Model-Base** (unfine-tuned), **Model-CE** (cross-entropy SFT), **Model-Sem** (DPO with deviation preferences). Identical initialization, corpus, optimizer, compute. A six-condition component-decomposed design (isolating deviation, provenance, and coherence contributions) is deferred to a follow-up experiment; the present three-condition design tests the composite hypothesis that the full bundle produces measurable improvement. If the composite succeeds, decomposition follows; if it fails, decomposition is moot.

5.4 Evaluation

Standard NLP benchmarks (MMLU, HellaSwag, ARC-Challenge, GSM8K) verify retained capability. Slop Composite Index (SCI): five component metrics (Net Deviation Signature, Cliché Frequency, Type-Token Ratio, N-gram Base-Rate Convergence, Surprise-Collapse Slope), each computed on 500 free-generation prompts under a fixed third-party reference model. $\text{SCI}(\theta) = \frac{1}{5}\sum z_i(\theta)$ with direction-corrected z-scores relative to Model-CE. SCI weights are uniform in the pre-registered primary analysis; a sensitivity analysis varying each by ±50% is reported as secondary. Pre-registered falsification: $\text{SCI}(\text{Model-Sem}) - \text{SCI}(\text{Model-CE}) > 0.25$.

Human preference evaluation: 500 prompt pairs × 3 raters (Prolific), blinded. 80% power for 56% preference rate at $\alpha = 0.05$ (binomial); a mixed-effects analysis (appropriate for nested rater data) is reported alongside.

5.5 Preference Validation Substudy

Before interpreting the DPO results, we validate the deviation-generated preference signal against human judgment. 100 randomly sampled preference pairs are independently rated by 3 human raters ("Which continuation do you prefer?"). If human agreement with the deviation-generated label is below 70%, the preference signal is

unreliable and the training intervention is compromised. This $50 substudy de-risks the entire intervention.

5.6 Judge Adversarial Validation

The frozen judge is validated against 200 adversarial strings (random tokens with citation markers). Mean π must be below 0.2. If the judge fails, training does not proceed (Perez et al. 2022).

5.7 Budget

Training (both architectures), judge fine-tuning, preference validation, evaluation suite, and human evaluation: **$3,000–$3,900**.

---

6. Anti-Goodhart Mechanism Design

A deviation-maximizing metric will be gamed. Manheim and Garrabrant (2019) taxonomize four varieties of Goodhart's law; Skalse et al. (2022) formalize reward hacking; Gao, Schulman, and Hilton (2023) demonstrate empirically that reward model overoptimization follows predictable scaling laws; Krakovna et al. (2020) document specification gaming across deployed systems. Six protections:

1. Entropy-floor capping (addresses extremal Goodhart). Texts scoring $\bar{\delta}$ above threshold must have conditional entropy exceeding $H_{\min} = 0.5$ bits. Blocks "surprise" from near-deterministic distributions.

2. Provenance-weighted damping (addresses adversarial Goodhart). Deviation weighted by π; high deviation with $\pi < 0.3$ damped toward zero.

3. Saturation threshold (addresses extremal Goodhart). Deviation saturates at the 95th percentile of a pre-registered 10,000-document OpenAlex calibration corpus.

4. Rolling-window variance penalty (addresses adversarial Goodhart). For F2, penalizes interventions whose deformation oscillates across intervals — blocks memetic-volatility farming.

5. Reference-model KL anchoring (inherited from DPO). The implicit KL penalty against θ_0 bounds distributional drift (Rafailov et al. 2023).

6. Black-box judge replacement test (addresses causal Goodhart). The frozen judge is replaced with a different architecture (same rubric) for a subset of evaluations. Spearman $\rho < 0.7$ triggers recalibration.

---

7. The Cheapest Dangerous Test

7.1 Setup

Corpora. GPT-wiki-intro (Bhat 2023): paired human/AI wiki introductions. HC3 (Guo et al. 2023): human/ChatGPT answer pairs. **Reference model.** `meta-llama/Llama-3.1-8B-Instruct`, frozen at the HuggingFace checkpoint as of deposit date. **Cost.** ~1 A100-hour for both corpora. **Pre-registration.** These predictions are pre-registered as a timestamped deposit on Zenodo prior to any computation (DOI to be inserted upon registration).

7.2 Pre-Registered Predictions

P1 (Machine-output convergence). AI-generated text in matched human/AI corpora exhibits statistically significant negative mean $\bar{\delta}$ relative to matched human text. Two-sided Mann-Whitney U at $\alpha = 0.05$, minimum effect size Cohen's $d > 0.5$. A positive result motivates a second-stage test on human-labeled low-quality AI text to determine whether the effect sharpens in the slop regime.

P2 (RLHF flattening). Post-RLHF chat-tuned models exhibit lower $\bar{\delta}$ than their pre-RLHF base counterparts on matched prompts — consistent with Ouyang et al.'s (2022) observation that RLHF produces more uniform outputs. Tested on `meta-llama/Llama-3.1-8B` (base) vs. `meta-llama/Llama-3.1-8B-Instruct` (chat-tuned), evaluated on 100 prompts from the OpenAssistant dataset matched by length bin. Limited to open-weight models where base weights are available.

P3 (Cross-judge consistency). The differential replicates under `mistralai/Mistral-7B-Instruct-v0.3`. Spearman rank correlation between per-output $\bar{\delta}$ rankings under Llama and Mistral exceeds 0.7. Failure indicates judge-specificity, not an intrinsic text property.

7.3 Outcome Logic

P1 failure disconfirms this paper's first high-stakes prediction — that benchmark AI text exhibits a negative signed-deviation signature. It would block the proposed training intervention in its current form while leaving the broader field-deformation measurement program open. P1 success with P3 failure retreats to a weaker, judge-relative claim. P1 and P3 success warrants the training intervention.

---

8. What This Paper Does Not Claim

1. That meaning is universally definable as deviation. The principle measures trajectory restructuring; aspects of meaning that do not produce distributional shift are outside its scope.
2. That the operationalizations are uniquely correct. F1 and F2 are canonical starting points.
3. That the anti-Goodhart machinery is sufficient against all gaming.

4. That cross-entropy training is wrong. It is insufficient for the target this framework specifies.
5. That $\mathcal{V}_T$ is ready for empirical use. It is not.
6. That the cheapest dangerous test will succeed. Failure is informative.
7. That this paper is independent of the Crimson Hexagonal Archive. It engages the founding formulation (Sharks 2026) directly and builds on companion protocols. What it claims is that a reader can evaluate the math, experiments, and predictions without engaging the broader institutional apparatus.

---

9. Roadmap

Horizon	Milestone	Budget
This week	Cheapest dangerous test (P1–P3) on GPT-wiki-intro + HC3, ~1 A100-hour	$50–$100
This month	Operationalization-stability: 50 texts evaluated under F1 across 3 reference models, cross-model $\bar{\delta}$ rank-correlation reported	$200–$500
This quarter	F2 protocol day-0 launch; 90-day window; 30 queries × 4 intervals × 6 surfaces	$1,500–$3,000
This quarter	Scale stability: P2 replicated across Llama-3.1 1B/8B/70B parameter family	$500–$1,000
This year	DPO training experiment (three conditions) + preference validation substudy	$3,000–$3,900
This year	Six-condition component decomposition (if three-condition composite succeeds)	$8,000–$12,000

Total: approximately $14,000–$19,000. Each major deposit reviewed by at least one external researcher in alignment, causal inference, computational linguistics, or information theory, selected for willingness to write damaging-if-warranted critiques.

References

Abadie, A. (2021). Using synthetic controls. *Journal of Economic Literature*, 59(2), 391–425.

Alemohammad, S., et al. (2023). Self-consuming generative models go MAD. *arXiv:2307.01850*.

Amari, S. (2016). *Information Geometry and Its Applications*. Springer.

Azar, M. G., et al. (2024). A general theoretical paradigm to understand learning from human feedback. *AISTATS 2024*.

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *British Journal for the Philosophy of Science*, 4(14), 147–157.

Bhat, S. (2023). GPT-wiki-intro. HuggingFace Datasets.

Bohnet, B., et al. (2022). Attributed question answering. *arXiv:2212.08037*.

Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS 2017*.

Deutsch, D., Doshi, R., & Roth, D. (2022). On the limitations of reference-free evaluations. *EMNLP 2022*.

Ethayarajh, K., et al. (2024). KTO: model alignment as prospect theoretic optimization. *arXiv:2402.01306*.

Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. *ICML 2023*.

Guo, B., et al. (2023). How close is ChatGPT to human experts? *arXiv:2301.07597*.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *NAACL 2001*, 159–166.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *ACL 2016*.

Hicks, D., et al. (2015). Bibliometrics: the Leiden Manifesto. *Nature*, 520, 429–431.

Holtzman, A., et al. (2020). The curious case of neural text degeneration. *ICLR 2020*.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Kirchenbauer, J., et al. (2023). A watermark for large language models. *ICML 2023*.

Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency, and nonequilibrium statistical physics. *Interface Focus*, 8(6).

Krakovna, V., et al. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind Technical Report.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Manheim, D., & Garrabrant, S. (2019). Categorizing variants of Goodhart's law. *arXiv:1803.04585*.

Meister, C., Cotterell, R., & Vieira, T. (2021). Revisiting the uniform information density hypothesis. *EMNLP 2021*.

Min, S., et al. (2023). FActScore: fine-grained atomic evaluation of factual precision. *EMNLP 2023*.

Mitchell, E., et al. (2023). DetectGPT: zero-shot machine-generated text detection using probability curvature. *ICML 2023*.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Perez, E., et al. (2022). Red teaming language models with language models. *EMNLP 2022*.

Pillutla, K., et al. (2021). MAUVE: measuring the gap between neural text and human text. *NeurIPS 2021*.

Rafailov, R., et al. (2023). Direct preference optimization. *NeurIPS 2023*.

Rashkin, H., et al. (2023). Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4).

Sadasivan, V. S., et al. (2023). Can AI-generated text be reliably detected? *arXiv:2303.11156*.

Schulman, J., et al. (2017). Proximal policy optimization algorithms. *arXiv:1707.06347*.

Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

Skalse, J., et al. (2022). Defining and characterizing reward hacking. *NeurIPS 2022*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

Stiennon, N., et al. (2020). Learning to summarize from human feedback. *NeurIPS 2020*.

Su, Y., et al. (2022). A contrastive framework for neural text generation. *NeurIPS 2022*.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391.

Welleck, S., et al. (2020). Neural text generation with unlikelihood training. *ICLR 2020*.

Ziegler, D. M., et al. (2019). Fine-tuning language models from human preferences. *arXiv:1909.08593*.

The founding formulation of the Semantic Deviation Principle is by Lee Sharks (DOI: 10.5281/zenodo.20250736). Companion protocols: 10.5281/zenodo.20251738, 10.5281/zenodo.20251740, 10.5281/zenodo.20251742.