# Orphaned Sophistication: Detecting AI-Generated Prose Through Structurally Unsupported Figurative Language

**Richard Quinn**
Independent Researcher
quinn.richard@gmail.com

## Abstract

We identify a novel stylometric artifact in large language model (LLM) prose generation: *orphaned sophistication*, figuratively sophisticated word choices that lack structural support from their surrounding context. Comparing 25 human-authored passages against 100 LLM-generated passages from three independent model families (Anthropic, OpenAI, Google), we demonstrate that LLMs produce polysemous words whose secondary semantic fields overlap with active figurative registers at rates significantly exceeding human prose (single-model, $n = 45$: Fisher's $p = 0.001$, Cohen's $h = 1.69$), without the metaphor chains, tonal preparation, or register consistency that characterize deliberate craft. We attribute this to training-weight distributional bias: literary texts exhibiting exceptional polysemous craft are disproportionately represented in training corpora, causing models to produce 99th-percentile figurative sophistication as default output. A three-dimensional orphanhood model (isolation, chain connectivity, tonal preparation) yields a deterministic detector achieving 28.0% true positive rate with 4% false positive rate (cross-family pooled, $n = 125$: Fisher's $p = 0.006$, Cohen's $h = 0.71$). The signal spans all three families: Anthropic (15–35%), OpenAI GPT-4o (15%), and Google Gemini 2.5 Flash (40%, $p = 0.004$, $h = 0.97$). Token probability probing confirms that flagged constructions are generated at elevated rates versus equivalent alternatives (e.g., 9.5× preference for personification vocabulary in Anthropic; 3.0× OpenAI; 2.0× Google). The uncanny valley of AI prose is a structural coherence failure, not a lexical quality failure, and it is measurable.

## 1 Introduction

The detection of AI-generated text has become a critical problem in computational linguistics, digital forensics, and publishing. Existing approaches fall broadly into two categories: statistical fingerprinting methods that measure distributional properties of token sequences (perplexity, burstiness, $n$-gram frequency profiles), and watermarking schemes that embed detectable signals during generation. Both share a fundamental limitation: they identify *that* a text is machine-generated without explaining *why* it reads as machine-generated. The qualitative experience of encountering AI prose, the uncanny valley sensation (Mori, 1970) that something is simultaneously competent and wrong, remains unformalized.

We present a third approach grounded in structural analysis of figurative language. Our central claim is that autoregressive language models, as a consequence of distributional biases in their training data, produce a specific and detectable artifact: figuratively sophisticated word choices that are structurally orphaned from the prose architecture that would justify them in human writing. A human author who writes "the hungry steel teeth" in a passage about a sawmill has, in deliberate literary prose, prepared that personification through tonal shifts, metaphor chains, or explicit signposting. An LLM produces the same construction as a default token prediction, without preparation, without continuation, and without architectural awareness that the construction requires either.

This paper makes four contributions: (1) empirical identification of the orphaned sophistication artifact through controlled experiments with formal statistical testing; (2) theoretical explanation through a training-weight over-indexing model; (3) a formal detection framework based on a three-dimensional orphanhood model, implemented as a fully deterministic rule-based algorithm; and (4) a semiotic interpretation connecting the signal to the distinction between Barthes's *signifiance* and *signification*.

# 2 Related Work

## 2.1 AI Text Detection

Current detection methods include perplexity-based classifiers (Mitchell et al., 2023), watermarking (Kirchenbauer et al., 2023), and supervised classifiers trained on LLM output distributions (Tian et al., 2023). These methods achieve variable accuracy and degrade across domains, paraphrasing attacks, and model updates (Krishna et al., 2023; Wu et al., 2025). Critically, none provides a structural explanation for *what* distinguishes AI prose from human prose at the level of craft.

We do not claim that orphaned sophistication detection replaces these methods. It operates in a different regime: short-form literary and descriptive prose where figurative language is expected.

## 2.2 Polysemy and Priming in LLM Output

Kugler (2025) demonstrates that LLM output exhibits a "flatter semantic space" than natural language (frequency-specificity correlation: $\rho \approx -0.3$ for LLMs vs. $\rho \approx -0.5$ to $-0.7$ for human text). This flat distribution is consistent with our over-indexing hypothesis. Jumelet et al. (2024) demonstrate that lexico-semantic overlap boosts token-level probability in transformers through structural priming effects, confirming the mechanistic foundation of our claim.

## 2.3 Coherence as a Latent Dimension

Shaib et al. (2025) develop a taxonomy of "AI slop" through expert annotation, finding that standard text metrics fail to capture coherence dimensions and that capable LLMs likewise fail to reliably identify slop. Our orphanhood framework provides one structural answer: it operationalizes a specific form of incoherence (figurative constructions arriving without architectural support) as a measurable, deterministic signal.

# 3 Theoretical Framework

## 3.1 Latent Semantic Recruitment

We define *Latent Semantic Recruitment* (LSR) as the phenomenon whereby an autoregressive language model, generating text within an active figurative register $R$, disproportionately selects polysemous words whose secondary semantic fields overlap with $R$.

Let $w$ be a word token with primary sense $s_1$ (contextually appropriate) and secondary sense $s_2$ (not contextually required). LSR occurs when $P(w \mid \text{context}, R) > P(w \mid \text{context}, \neg R)$ specifically because the embedding of $s_2$ overlaps with $R$ in the model's representation space.

This follows from the standard transformer output computation (Vaswani et al., 2017). The logit for token $w$ at position $t$ is $z_w = \mathbf{v}_w^\top \mathbf{h}_t$, where $\mathbf{v}_w$ is the output embedding and $\mathbf{h}_t$ is the hidden state. When context contains register-activating content, $\mathbf{h}_t$ encodes semantic components overlapping with register-aligned secondary senses. For polysemous words, $\mathbf{v}_w$ encodes both $s_1$ and $s_2$, and the inner product with a register-active $\mathbf{h}_t$ is elevated compared to a monosemous alternative encoding only $s_1$.

## 3.2 Training-Weight Over-Indexing

LSR explains the mechanism, but not why the result is detectable. Human writers also select polysemous words. The critical question is why LLM polysemous usage is distinguishable.

We propose the **training-weight over-indexing hypothesis**: training corpora contain a distributional bias that systematically over-represents exceptional figurative prose. The texts exhibiting the most sophisticated polysemous craft (Conrad, McCarthy, Woolf, Morrison) receive the most analytical attention, pedagogical citation, and anthology inclusion, producing the most duplication across training data. Under standard cross-entropy training, tokens appearing more frequently contribute proportionally more to the cumulative gradient. The model therefore learns to reproduce this level of sophistication not as exceptional but as the expected register of competent prose.

The result is a distributional inversion. In the population of human writers, polysemous craft at the level of Conrad or McCarthy occupies the far right tail. In the model's learned distribution, it occupies the mode. We hypothesize that when this disparity occurs, it constitutes a detectable signal.

**Caveat.** A direct demonstration would require measuring the frequency of specific figurative constructions in training data and correlating that frequency with generation probability, an analysis requiring training-data access we do not have. The hypothesis is argued from distributional logic, consistent with the observed signal, but not independently verified.

### 3.3 Orphaned Sophistication

The over-indexing hypothesis predicts that LLMs will produce sophisticated figurative language *without the structural architecture that earns it*. We define **orphaned sophistication** as a figurative construction satisfying three conditions:

**Isolation.** The figurative density of the sentence containing $w$ is significantly higher than its neighbors (window $\pm 2$ sentences). Score: $\min(1.0, (\varphi(s_w) - \bar{\varphi}(N(s_w)))/\tau_1)$, where $\tau_1 = 0.2$.

**Chain disconnection.** The register field activated by $s_2$ is not activated by other words within $\pm 3$ sentences. In human literary prose, figurative constructions participate in metaphor chains. Score: 0 connections = 1.0, 1 = 0.6, 2 = 0.2, 3+ = 0.0.

**Lack of preparation.** The context is scored for signposting markers: simile constructions, explicit frame-setting, tonal shifts (sentence-length ratio $>$ 2.5:1), and figurative density in adjacent sentences ($> 0.15$). Score ranges from 1.0 (fully unprepared) to 0.0 (fully prepared).

A word's orphanhood score is the arithmetic mean of its three test scores. A word exceeding 0.6 is classified as *orphaned*; words below that threshold are classified as *structurally integrated* (the figurative construction participates in the surrounding prose architecture). The threshold was set a priori, not optimized on test data.

### 3.4 Semiotic Interpretation

The framework admits an interpretation through Barthes's semiotic theory (Barthes, 1970, 1973), though we acknowledge this narrows Barthes's framework considerably. In *S/Z*, *signifiance* concerns the plurality of meaning generated by the interaction of multiple codes in writerly texts; our usage maps a more architectural reading onto the term, treating *signifiance* as requiring structural scaffolding. This narrowing is deliberate and operational.

When Conrad writes "the rudder would bite," the word performs something closer to *signifiance*: it participates in a novel-length architecture. When an LLM writes "the hungry steel teeth," the same semantic content is present but the structural labor is absent. The word performs *signification* without *signifiance*. Our three tests map onto this distinction: isolation measures sustained vs. anomalous sophistication; chain connectivity measures productive labor vs. standalone activation; preparation measures deliberate register transition vs. its absence.

A necessary caveat: a sufficiently long LLM text may, through stochastic density alone, produce passages scoring well on all three dimensions. Our detector measures necessary conditions for *signifiance* (structure is present) but not the sufficient condition (structure was produced through authorial labor). This is why we describe the framework as identifying *orphaned* sophistication rather than *unearned* sophistication.

## 4 Experimental Method

### 4.1 Corpus Construction

We assembled three corpora across five physical-register domains (ocean storm, kitchen/restaurant, blacksmith/forge, battlefield surgery, sawmill/logging):

**Human Corpus A (Published).** 20 passages (~100–200 words), drawn from published fiction and nonfiction spanning 1902–2016, four per domain. Authors include Conrad (1902), Hemingway (1952), McCarthy (1985), Bourdain (2000), Proulx (2016), Powers (2012), Barker (1991), Remarque (1929), Orwell (1933), and ten others.[1]

**Human Corpus B (Non-professional).** 5 passages (~150–250 words), hand-written by a non-professional writer under experimental conditions (one per domain, written under time pressure without revision, before the detection framework was developed).

**LLM Corpora C.** 100 passages total from five model runs spanning three independent families: Claude Sonnet 4 (20 + 20 replication, Anthropic), Claude Haiku 3.5 (20, Anthropic), GPT-4o (20, OpenAI), Gemini 2.5 Flash (20, Google). All generated via API at temperature 1.0, four passages per domain, identical prompts across models. Prompts requested 150–200 word passages specifying physical detail, past tense, third person, no dialogue.

### 4.2 Corpus Provenance

Corpus A passages were reproduced from LLM training data, introducing a potential circularity (Section 7.6). Corpus B provides an uncontaminated baseline. The detection instrument is deterministic (no LLM judgment in scoring), so circularity applies only to corpus construction.

[1] Full corpus: O'Brian (1969), Junger (1997), Buford (2006), Fisher (1954), Thompson (1945), McPhee (1975), Sturt (1923), Hooker (1968), Kesey (1964), Berry (2000), Pollan (1997).

### 4.3 Detection Instrument

We developed three successive detection instruments. Detector v1 (rate-based) counted polysemous words with register-aligned secondary senses; it was discarded because human and LLM rates were too similar. Detector v2 introduced domain-literal filtering, personification detection, and metaphor signpost detection, achieving strong separation but unable to distinguish skilled human figurative construction from LLM-generated equivalents at the individual word level.

The reported instrument (detector v3) identifies figurative polysemous words using v2's mechanisms, then subjects each candidate to the three orphanhood tests defined in Section 3.3. The algorithm processes each sentence, identifying words that (a) are not in the domain-literal set for the passage's domain (34–41 words per domain), (b) appear in at least one of six register fields (consumption, personification, body, violence, fire/heat, water/weather; 16–29 words each), and (c) exhibit figurative usage (personification, animate verb, or animate-quality modifier). Qualifying words receive the three orphanhood scores; the arithmetic mean must exceed 0.6 for classification as orphaned.

The detector is fully deterministic, requiring neither neural networks, LLM judgment, nor learned parameters. All thresholds are set a priori. Domain-literal filtering is conservative by design: it suppresses only words whose primary sense denotes the domain activity. A word like "grip" is not added to the blacksmith domain-literal set even though blacksmiths literally grip tools, because its primary sense (physical grasping) is not specific to blacksmithing (see supplementary material E.4). Domain-literal filtering operates on the passage's declared domain, assigned at corpus construction time.[2]

### 4.4 Statistical Methods

All comparisons use Fisher's exact test (appropriate for small-sample count data). We report one-sided $p$-values (testing the directional hypothesis that LLM rates exceed human rates) and two-sided $p$-values. Confidence intervals use the Clopper-Pearson exact method ($\alpha = 0.05$). Effect sizes are Cohen's $h$, where $h > 0.8$ is conventionally large. No multiple-comparison correction is applied to the primary analysis (single pre-specified comparison); per-domain exploratory analyses are flagged as uncorrected.

| Source | $n$ | Orphaned | Flagged | Rate |
|---|---|---|---|---|
| Published human | 20 | 1 | 1/20 | 5.0% |
| Non-prof. human | 5 | 0 | 0/5 | 0% |
| **All human** | **25** | **1** | **1/25** | **4.0%** |
| Sonnet (primary) | 20 | 9 | 7/20 | 35.0% |

Table 1: Detector v3 results, primary experiment. Fisher's $p = 0.010$ (one-sided), Cohen's $h = 0.86$.

### 4.5 Experimental Design

The primary experiment (Experiment 8c/v3): all 25 human and 20 LLM passages (Corpus C-Sonnet) were processed by detector v3; orphanhood scores were computed for each flagged word; results were aggregated by source and domain. Cross-model replication used 20 additional passages each from Sonnet 4 and Haiku 3.5 under identical conditions. Cross-family validation used 20 passages each from GPT-4o and Gemini 2.5 Flash, testing whether the signal generalizes beyond a single model family.

## 5 Results

### 5.1 Primary Analysis

The single human detection was Conrad's "the rudder would bite again," a nautical usage where "bite" is arguably domain-literal. In the 5 non-professional passages, zero detections occurred.

The detector v2 analysis (unjustified figurative polysemy, before orphanhood filtering) showed a rate ratio of 18.8× (LLM 0.750 per passage vs. human 0.040), Fisher's $p = 0.001$, Cohen's $h = 1.69$. The v3 orphanhood model correctly reclassified 6 of the 15 v2 detections as "integrated": cases where the model had accidentally produced chain connectivity (e.g., "bit" appearing near "teeth" or "hungry").

### 5.2 Qualitative Analysis

The most striking LLM passage was L06 (sawmill domain, Sonnet), which contained four orphaned words:

By contrast, human passages employing figurative language do so within explicitly prepared frames. The non-professional sawmill passage ("Life shaves pieces of your health off… Bertha takes that, too") signposts the saw-as-life metaphor

[2] Full algorithmic specification, domain-literal sets, register field taxonomy, and all code will be released upon publication.

| Word | Register | Score | Iso | Chain | Prep |
|---|---|---|---|---|---|
| “hungry” | personif. | 0.88 | 0.6 | 1.0 | 1.0 |
| “stubborn” | personif. | 0.73 | 0.2 | 1.0 | 1.0 |
| “bite” | consumption | 0.73 | 0.2 | 1.0 | 1.0 |
| “roar” | water/weath. | 0.80 | 0.4 | 1.0 | 1.0 |

Table 2: Four orphaned words in passage L06 (sawmill, Sonnet). Four distinct register fields, no chain connectivity between any, no preparation for any.

| Model | Family | $n$ | Rate | $h$ | $p$ |
|---|---|---|---|---|---|
| Human | – | 25 | 4.0% | – | – |
| Gemini 2.5 | Google | 20 | 40.0% | 0.97 | .004 |
| Sonnet (orig) | Anth. | 20 | 35.0% | 0.86 | .010 |
| Haiku 3.5 | Anth. | 20 | 35.0% | 0.86 | .010 |
| GPT-4o | OpenAI | 20 | 15.0% | 0.39 | .224 |
| Sonnet (repl) | Anth. | 20 | 15.0% | 0.39 | .224 |
| **All LLM** | **3 fam.** | **100** | **28.0%** | **0.71** | **.006** |

Table 3: Cross-family validation. All $p$-values one-sided Fisher’s exact. Pooled 95% CI: LLM [0.195, 0.379]; human [0.001, 0.204]. Power: 93.8%.

(“That’s life, that is”), develops it across multiple sentences, and connects to a chain of related vocabulary. The detector correctly classifies this as structurally integrated.

### 5.3 Cross-Model and Cross-Family Validation

Three of five model runs are individually significant. The pooled analysis across three independent families provides the definitive test ($p = 0.006$, $h = 0.71$, power 93.8%).

**Cross-family word convergence.** The same register fields and often the same words (“roar,” “hungry,” “angry,” “bit”) recur across independently trained models from three organizations, strongly supporting the over-indexing hypothesis.

**Gemini signal strength.** Gemini produced the highest orphanhood rate (40%, 8/20), with one blacksmith passage producing three orphaned words. The word “grip” in that passage warrants scrutiny as potentially domain-literal (blacksmiths literally grip tools); removing it would reduce orphaned words from 10 to 9 without affecting the passage-level rate.

### 5.4 Per-Domain Distribution

The kitchen domain produced a null result (0/5 human, 0/4 LLM), the only domain with zero detections. Culinary language is inherently action-oriented and consumption-related, so words that would register as figurative in other domains are domain-literal in a kitchen context. The detector correctly identifies these as non-figurative.

| Domain | Human | LLM | Fisher $p$ |
|---|---|---|---|
| Sawmill | 0/5 | 4/4 | 0.008* |
| Surgery | 0/5 | 4/4 | 0.008* |
| Blacksmith | 0/5 | 1/4 | 0.444 |
| Ocean storm | 0/5 | 1/4 | 0.444 |
| Kitchen | 0/5 | 0/4 | 1.000 |

Table 4: Per-domain results (original Sonnet run). *Uncorrected for multiple comparisons.

| Config | Dims | Hum | LLM | $h$ | $p$ |
|---|---|---|---|---|---|
| Full | I+C+P | 1/25 | 27/100[†] | .690 | .008 |
| −Isolation | C+P | 1/25 | 33/100 | .821 | .002 |
| −Chain | I+P | 1/25 | 21/100 | .549 | .035 |
| −Preparation | I+C | 2/25 | 26/100 | .497 | .041 |

Table 5: Ablation study (125 passages). I = isolation, C = chain, P = preparation. [†]One Sonnet passage falls at the classification boundary (score $\approx 0.60$), producing 27/100 in the unified ablation pass vs. 28/100 in the incremental main analysis. The difference is due to floating-point variation and does not affect qualitative conclusions.

### 5.5 Ablation Study

The results reveal an asymmetric architecture. Removing chain connectivity reduces $h$ by 0.141, confirming that chain detection captures discriminative signal. Removing preparation produces the largest $h$ degradation (0.193) and doubles the human false positive rate from 4.0% to 8.0%, indicating that preparation is the dimension most responsible for specificity.

Removing isolation *increases* $h$ to 0.821 because isolation functions as a conservative filter, suppressing true positives where the figurative spike coincides with mildly elevated neighborhood density. All four configurations maintain significance ($p < 0.05$).

## 6 Mechanism Validation

### 6.1 Monte Carlo Logit Proxy

A Monte Carlo simulation generated 100,000 random word-context pairings and computed register-overlap scores. Of 15 register-field/domain pairings, 12 showed zero overlap between the random distribution and the observed LLM orphan scores ($p < 10^{-5}$ each), confirming that the observed scores are not achievable by chance co-occurrence.

| Condition | Orphan score |
|---|---|
| Suppressed (“avoid figurative”) | 0.000 |
| Neutral (no instruction) | 0.098 |
| Amplified (“use vivid language”) | 0.755 |

Table 6: Dose-response experiment. Orphanhood is register-dependent and dose-responsive.

| Probe | Register | Anth. | OAI | Gem. |
|---|---|---|---|---|
| SAW_BITE | Consumption | 1.41 | 3.43 | $\infty$* |
| OCEAN_ROAR | Vocalization | 1.17 | 1.44 | 3.00 |
| FORGE_STUB | Personif. | 9.50 | 3.00 | 2.00 |
| SURG_SCRM | Vocalization | 1.38 | 2.50 | $\infty$* |

Table 7: Preference ratios (literary/equivalent) for the four probes showing consistent literary preference across all three families. *Infinite: literary words present, zero equivalents in all completions. Equivalent word lists are not exhaustive; $\infty$ indicates strong directional preference, not absence of all non-literary vocabulary. Four additional probes showed mixed/reversed preferences.

### 6.2 Dose-Response

### 6.3 Token Probability Probing

To test the over-indexing hypothesis at the generation level, we designed eight probes targeting constructions the detector most frequently flags. Each probe provides a physical-register context (e.g., “Write a paragraph describing a sawmill blade cutting through hardwood”) and generates $N$ completions at temperature 1.0 ($N$=20 for Anthropic/Gemini, $N$=10 for OpenAI). For each completion, we count “literary” words (high-prestige figurative constructions: “bite,” “teeth,” “hungry,” “stubborn,” etc.) versus semantically equivalent alternatives (“cut,” “slice,” “hard,” “rigid,” etc.). One probe (KITCHEN_ALIVE) tests anthropomorphic vitality constructions in a domain where the primary detection mechanism does not apply, since kitchen vocabulary is inherently consumption-register and therefore domain-literal. Full word lists are provided with the code release.

The FORGE_STUBBORN probe produced the strongest signal: Anthropic generated personification vocabulary at 9.5× the rate of physical-property alternatives. All three families showed aggregate literary preference (Anthropic 1.30×, OpenAI 1.45×, Gemini 1.91×).

## 7 Discussion

### 7.1 The Uncanny Valley Formalized

The orphaned sophistication framework provides a structural account of the “uncanny valley” of AI prose (Mori, 1970). The deficiency lies not in vocabulary or grammar but in the *relationship between sophistication and structure*: the text produces figurative constructions implying architectural control, but the architecture is absent. This formalizes the observation that AI prose reads as “too good” at the sentence level while failing at the paragraph level (Shaib et al., 2025).

### 7.2 Why This Is Not Watermarking

Orphaned sophistication is not an imposed signal; under the over-indexing hypothesis, it is an emergent artifact of training. If correct, the signal would resist removal by post-processing or prompt engineering. Whether fine-tuning could selectively reduce orphanhood without degrading prose quality is an open empirical question.

### 7.3 Alternative Explanations

**Attention span.** LLMs may produce orphaned sophistication due to attention-window limitations rather than training-weight bias. However, this does not explain why the *specific* constructions are so consistent across independent generations. Attention limitations would predict random figurative orphanhood; we observe patterned orphanhood.

**Mode collapse.** All passages were generated at temperature 1.0, a regime that Holtzman et al. (2020) demonstrate substantially reduces repetitive degeneration. We observe the same figurative *strategy* expressed in varied syntactic frames, more consistent with a learned register preference. Temperature 1.0 reduces repetition while preserving the probability distribution’s shape, precisely the regime where over-indexing effects would manifest as preferential selection.

### 7.4 Implications

If the over-indexing hypothesis is correct, orphaned sophistication should be present in all LLMs trained on standard web corpora. The signal should persist across architectures because it arises from distributional properties of training data. The signal is interpretable: a detection report points to specific words, explains why they are orphaned, and provides structural explanation. For writers using LLMs collaboratively, the framework provides actionable revision guidance: flagged passages require not deletion but *architecture* (build a chain, prepare the register shift, sustain figurative density).

### 7.5 The Generalizable Principle

The hypothesis predicts that any domain where models are trained on corpora dominated by exceptional exemplars will exhibit an analogous artifact. Du et al. (2026) independently identify “hard negatives” in computational drug repurposing: well-studied compounds appearing ideal due to high knowledge-graph connectivity but failing clinically. The mechanism is structurally identical: FDA-approved drugs dominate training corpora through citation and patent literature, producing the same over-indexing dynamic. Graph neural networks learn to produce binding moieties resembling successful drugs as default output, locally brilliant binding predictions structurally orphaned from the ADMET properties that would make them clinically viable. This corresponds to the “activity cliff” problem in medicinal chemistry (Stumpfe and Bajorath, 2012).

The mapping is exact: a high-affinity binding moiety without metabolic stability is an orphaned figurative word. Isolation, chain disconnection, and lack of preparation all have molecular analogues. The convergence, observed independently in a different modality by researchers with no knowledge of the orphanhood framework, suggests the pathology is a general property of training-data distributional bias. (We note that Du et al. is a bioRxiv preprint not yet peer-reviewed.) See Section 7.6, item 8, for the corresponding research agenda.

### 7.6 Future Directions

We identify nine directions for future work, ordered from immediate empirical extensions to more speculative theoretical predictions.

1. **Scaled replication.** A study with $\geq$200 passages across additional domains would strengthen per-model and per-domain claims and permit meaningful multiple-comparison correction.

2. **Baseline detector comparison.** Running DetectGPT, GPTZero, and watermark-based methods on the same corpus would establish whether the orphanhood signal is orthogonal to existing detection methods.

3. **Open-weight models.** Testing on Llama 3, Mistral, and Gemma would extend the generalizability claim and permit direct inspection of training data and model weights.

4. **Mid-tier human prose.** Characterizing the false positive rate on competent-but-not-elite writing (workshop fiction, genre fiction, journalistic features) is essential before any practical deployment.

5. **Uncontaminated human corpus.** Passages directly transcribed from physical books would eliminate the provenance concern described in Section 4.2.

6. **Abstract registers.** Extension to emotional, philosophical, and political registers requires new domain-literal sets. Whether the signal persists where figurative language is less physically grounded is an open question.

7. **Fine-tuning resistance.** If the over-indexing hypothesis is correct, the signal should resist removal by post-processing. Whether targeted fine-tuning could selectively reduce orphanhood without degrading prose quality would constitute a strong test: if the signal can be trained out cheaply, the over-indexing explanation is weakened.

8. **Cross-domain validation.** Systematic testing in legal document generation, scientific writing, and financial analysis would determine whether the three-dimensional diagnostic transfers with appropriate domain vocabulary.

9. **Constrained-vocabulary environments.** The over-indexing mechanism predicts a qualitatively different outcome in domains where polysemy is structurally impossible. In computer code, every token is monosemous: a function name can only mean what its definition specifies. The same training-weight concentration that produces orphaned sophistication in prose would, in code, produce pattern-matching on well-architected exemplars from heavily-cited repositories (Chen et al., 2021; Austin et al., 2021). Because code has deterministic correctness criteria, the compiler enforces structural integration: a locally sophisticated pattern either works within the surrounding architecture or fails

to execute. The over-indexing liability in literary prose (locally brilliant, structurally orphaned) becomes, we hypothesize, an asset in code (locally idiomatic, structurally constrained to integrate). This predicts that models exhibiting the strongest over-indexing should perform disproportionately well in constrained-vocabulary environments. Testing would require adapting the orphanhood framework's logic to code-structure analogues (function isolation from calling context, API chain connectivity, type-preparation through declarations) and comparing over-indexing signatures across model families in both prose and code tasks.

## Limitations

1. **Sample size.** $n = 125$ (25 human, 100 LLM) across 5 domains and 3 families. Statistically significant but modest for strong generalization.

2. **Human corpus provenance.** Corpus A was reproduced from LLM training data. Corpus B provides an uncontaminated baseline, but ideal replication would use passages transcribed from physical books.

3. **Human corpus skill ceiling.** The detector's FPR was measured against elite literary prose. Its behavior on mid-tier prose (workshop fiction, genre fiction, journalism) is unknown and is a priority for future validation.

4. **No baseline comparison.** We have not run existing detectors (DetectGPT, GPTZero) on the same corpus.

5. **Domain specificity.** Currently implemented for five physical-register domains. Extension to abstract registers requires additional word sets.

6. **Passage length.** Orphanhood tests operate on passage-length windows. Earned figurative language can score as orphaned if excerpted from longer works.

7. **Same-model variance.** The Sonnet replication (15%) was individually non-significant ($p = 0.224$). Post-hoc power analysis shows 80% power to detect $h \geq 0.53$ at $n = 20$; the replication's $h = 0.39$ falls below this threshold. The non-significant result reflects power limitations, not absent signal.

## 8 Conclusion

We have identified and formalized a novel artifact of autoregressive language generation: orphaned sophistication. Our experiments provide statistically significant evidence across three independent model families (Fisher's $p = 0.006$, Cohen's $h = 0.71$, $n = 125$). The artifact arises, we argue, from training-weight over-indexing on exceptional exemplars, causing models to produce locally sophisticated outputs without the structural architecture that would earn them.

The detection signal is structural and interpretable: it provides not merely a classification but a diagnosis of where the architecture is missing and what kind of work would repair it. The core contribution is a reframing: the uncanny valley of AI prose is a structural coherence failure, not a lexical quality failure, and it is measurable. The machine does not write badly. It writes too well, in moments that have not been earned.

## LLM Assistance Disclosure

This paper was written with the assistance of several large language models used as research tools. Claude Opus 4 (Anthropic) was used to implement detection algorithms, reproduce published passages from training data for Corpus A, generate statistical analyses, and draft the manuscript under the author's direction. Claude Sonnet 4, Claude Haiku 3.5, GPT-4o, and Gemini 2.5 Flash generated the respective LLM test corpora. No model involved in corpus generation was involved in scoring: all detection was performed by deterministic rule-based algorithms.

## References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *arXiv*. ArXiv:2108.07732.

Pat Barker. 1991. *Regeneration*. Viking.

Roland Barthes. 1970. *S/Z*. Éditions du Seuil.

Roland Barthes. 1973. *Le Plaisir du texte*. Éditions du Seuil.

Wendell Berry. 2000. *Jayber Crow*. Counterpoint.

Anthony Bourdain. 2000. *Kitchen Confidential*. Bloomsbury.

Bill Buford. 2006. *Heat*. Knopf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv*. ArXiv:2107.03374.

Joseph Conrad. 1902. *Typhoon*. Heinemann.

Ruiqi Du, Maximilian Fung, Yifei Hu, and David Liu. 2026. Overcoming topology bias and cold-start limitations in drug repurposing: A clinical-outcome-aligned LLM framework. *bioRxiv*.

M.F.K. Fisher. 1954. *The Art of Eating*. World Publishing.

Ernest Hemingway. 1952. *The Old Man and the Sea*. Scribner.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR 2020*.

Richard Hooker. 1968. *MASH: A Novel About Three Army Doctors*. Morrow.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Syntactic structural priming in large language models. In *Proceedings of ACL 2024*.

Sebastian Junger. 1997. *The Perfect Storm*. Norton.

Ken Kesey. 1964. *Sometimes a Great Notion*. Viking.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of ICML 2023*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Proceedings of NeurIPS 2023*.

R. Kugler. 2025. Polysemy patterns in large language model output. *arXiv*. ArXiv:2511.21334.

Cormac McCarthy. 1985. *Blood Meridian*. Random House.

John McPhee. 1975. *The Survival of the Bark Canoe*. Farrar, Straus and Giroux.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of ICML 2023*.

Masahiro Mori. 1970. The uncanny valley. *Energy*, 7(4):33–35. K. F. MacDorman & N. Kageki, Trans., IEEE Robotics & Automation Magazine, 19(2), 2012.

Patrick O'Brian. 1969. *Master and Commander*. Collins.

George Orwell. 1933. *Down and Out in Paris and London*. Gollancz.

Michael Pollan. 1997. *A Place of My Own*. Random House.

Kevin Powers. 2012. *The Yellow Birds*. Little, Brown.

Annie Proulx. 2016. *Barkskins*. Scribner.

Erich Maria Remarque. 1929. *Im Westen nichts Neues [All Quiet on the Western Front]*. Propyläen.

Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C. Wallace. 2025. Detection and measurement of AI-generated text quality dimensions: Expert taxonomy and span-level annotation. *arXiv*. ArXiv:2509.19163.

Dagmar Stumpfe and Jürgen Bajorath. 2012. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7):2932–2942.

George Sturt. 1923. *The Wheelwright's Shop*. Cambridge University Press.

Flora Thompson. 1945. *Lark Rise to Candleford*. Oxford University Press.

Edward Tian and 1 others. 2023. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. Preprint.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.