

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

Anonymous TACL submission

Abstract

We identify a novel stylometric artifact in large language model (LLM) prose generation: *orphaned sophistication*, the production of figuratively sophisticated word choices that lack structural support from their surrounding context. Through controlled experiments comparing 25 human-authored passages against 100 LLM-generated passages from five model runs spanning 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 (initial single-model analysis, $n = 45$: Fisher’s exact test, $p = 0.001$, Cohen’s $h = 1.69$). We propose a theoretical explanation rooted in training-weight distributional bias and formalize a three-dimensional orphanhood model (isolation, chain connectivity, tonal preparation), implementing a deterministic rule-based detector achieving 28.0% true positive rate on LLM prose with 4% false positive rate on human prose (cross-family pooled analysis, $n = 125$: Fisher’s $p = 0.006$, Cohen’s $h = 0.71$). The signal spans all three families tested: Anthropic (15–35%), OpenAI GPT-4o (15%), and Google Gemini 2.5 Flash (40%, $p = 0.004$). Token probability probing confirms that the specific constructions the detector flags are generated at elevated rates compared to semantically equivalent alternatives across all three families (e.g., 9.5 \times preference for personification vocabulary in Anthropic models; 3.0 \times OpenAI; 2.0 \times Google). The central finding is that the uncanny valley of AI prose is a structural coherence failure, not a lexical quality failure, and it is measurable. We provide a semiotic interpretation grounding the signal in the distinction between Barthes’s *significance* and *signification*, and identify a structurally

identical pathology in computational drug repurposing, suggesting domain generality.

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 *significance* 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 $P(w | \text{context}, R) > P(w | \text{context}, \neg R)$ 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 ± 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 ± 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*, *significance* 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 *significance* as requiring structural scaffolding. This narrowing is deliberate and operational.

When Conrad writes “the rudder would bite,” the word performs something closer to *significance*: 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 *significance*. 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 *significance* (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

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

300 requested 150–200 word passages specifying phys-
 301 ical detail, past tense, third person, no dialogue.
 302

303 4.2 Corpus Provenance

304 Corpus A passages were reproduced from LLM
 305 training data, introducing a potential circularity
 306 (Section 7.6). Corpus B provides an uncontami-
 307 nated baseline. The detection instrument is deter-
 308 ministic (no LLM judgment in scoring), so circu-
 309 larity applies only to corpus construction.

310 4.3 Detection Instrument

311 We developed three successive detection instru-
 312 ments. Detector v1 (rate-based) counted poly-
 313 semous words with register-aligned secondary
 314 senses; it was discarded because human and LLM
 315 rates were too similar. Detector v2 introduced
 316 domain-literal filtering, personification detection,
 317 and metaphor signpost detection, achieving strong
 318 separation but unable to distinguish skilled human
 319 figurative construction from LLM-generated equiv-
 320 alents at the individual word level.

321 The reported instrument (detector v3) identifies
 322 figurative polysemous words using v2’s mech-
 323 anisms, then subjects each candidate to the three
 324 orphanhood tests defined in Section 3.3. The algo-
 325 rithm processes each sentence, identifying words
 326 that (a) are not in the domain-literal set for the
 327 passage’s domain (34–41 words per domain), (b)
 328 appear in at least one of six register fields (con-
 329 sumption, personification, body, violence, fire/heat,
 330 water/weather; 16–29 words each), and (c) exhibit
 331 figurative usage (personification, animate verb, or
 332 animate-quality modifier). Qualifying words re-
 333 ceive the three orphanhood scores; the arithmetic
 334 mean must exceed 0.6 for classification as or-
 335 phaned.

336 The detector is fully deterministic, requiring nei-
 337 ther neural networks, LLM judgment, nor learned
 338 parameters. All thresholds are set a priori. Domain-
 339 literal filtering is conservative by design: it sup-
 340 presses only words whose primary sense denotes
 341 the domain activity. A word like “grip” is not
 342 added to the blacksmith domain-literal set even
 343 though blacksmiths literally grip tools, because its
 344 primary sense (physical grasping) is not specific
 345 to blacksmithing (see supplementary material E.4).
 346 Domain-literal filtering operates on the passage’s
 347 declared domain, assigned at corpus construction
 348 time.²

349 ²Full algorithmic specification, domain-literal sets, register

Source	<i>n</i>	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%

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

358 4.4 Statistical Methods

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

370 4.5 Experimental Design

371 The primary experiment (Experiment 8c/v3): all 25
 372 human and 20 LLM passages (Corpus C-Sonnet)
 373 were processed by detector v3; orphanhood scores
 374 were computed for each flagged word; results were
 375 aggregated by source and domain. Cross-model
 376 replication used 20 additional passages each from
 377 Sonnet 4 and Haiku 3.5 under identical condi-
 378 tions. Cross-family validation used 20 passages
 379 each from GPT-4o and Gemini 2.5 Flash, test-
 380 ing whether the signal generalizes beyond a single
 381 model family.

382 5 Results

383 5.1 Primary Analysis

384 The single human detection was Conrad’s “the
 385 rudder would bite again,” a nautical usage where
 386 “bite” is arguably domain-literal. In the 5 non-
 387 professional passages, zero detections occurred.

388 The detector v2 analysis (unjustified figurative
 389 polysemy, before orphanhood filtering) showed a
 390 rate ratio of $18.8 \times$ (LLM 0.750 per passage vs. hu-
 391 man 0.040), Fisher’s $p = 0.001$, Cohen’s $h = 1.69$.
 392 The v3 orphanhood model correctly reclassified
 393 6 of the 15 v2 detections as “integrated”: cases
 394 where the model had accidentally produced chain

395 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%.

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 (“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

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 ≈ 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.

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.

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 den-

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.

sity. 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.

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 $9.5\times$ the rate of physical-property alternatives. All three families showed aggregate literary preference (Anthropic $1.30\times$, OpenAI $1.45\times$, Gemini $1.91\times$).

Probe	Register	Anth.	OAI	Gem.
SAW_BITE	Consumption	1.41	3.43	∞^*
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	∞^*

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; ∞ indicates strong directional preference, not absence of all non-literary vocabulary. Four additional probes showed mixed/reversed preferences.

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. Tem-

perature 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 9, 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 ≥ 200 passages across additional domains would strengthen per-model and per-domain claims and permit meaningful multiple-comparison correction. 650
651
652
653
654
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. 655
656
657
658
659
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. 660
661
662
663
664
665
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. 666
667
668
669
5. **Uncontaminated human corpus.** Passages directly transcribed from physical books would eliminate the provenance concern described in Section 4.2. 670
671
672
673
674
675
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. 676
677
678
679
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. 680
681
682
683
684
685
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. 686
687
688
689
690
691
692
693

700	9. Constrained-vocabulary environments.	750
701	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 (??). 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.	751
702		752
703		753
704		754
705		755
706		756
707		757
708		758
709		759
710		760
711		761
712		762
713		763
714		764
715		765
716		766
717		767
718		768
719		769
720		770
721		771
722		772
723		773
724		774
725		775
726		776
727		777
728		778
729		779
730	8 Conclusion	780
731	We have identified and formalized a novel artifact	781
732	of autoregressive language generation: orphaned	782
733	sophistication. The artifact arises from training-	783
734	weight over-indexing on exceptional exemplars,	784
735	causing models to produce locally sophisticated	785
736	outputs without the structural integrity that would	786
737	justify them. In pooled analysis across three inde-	787
738	pendent model families, our detector achieves sig-	788
739	nificant separation between human and LLM text	789
740	(Fisher’s $p = 0.006$, Cohen’s $h = 0.71$, $n = 125$).	790
741	The uncanny valley of AI prose is a measurable	791
742	structural coherence failure. The detector identi-	792
743	fies where figurative constructions lack architec-	793
744	tural support and what structural work would close	794
745	the gap. Whether the signal persists under fine-	795
746	tuning, extends to abstract registers, or inverts in	796
747	constrained-vocabulary environments where com-	797
748	pliers enforce integration remains to be tested (Sec-	798
749	tion 7.6).	799
750	Limitations	788
751	1. Sample size. $n = 125$ (25 human, 100 LLM)	789
752	across 5 domains and 3 families. Statistically	790
753	significant but modest for strong generaliza-	791
754	tion.	792
755	2. Human corpus provenance. Corpus A was	793
756	reproduced from LLM training data. Corpus B provides an uncontaminated baseline,	794
757	but ideal replication would use passages transcribed from physical books.	795
758	3. Human corpus skill ceiling. The detector’s	796
759	FPR was measured against elite literary prose.	797
760	Its behavior on mid-tier prose (workshop fiction,	798
761	genre fiction, journalism) is unknown	799
762	and is a priority for future validation.	799
763	4. No baseline comparison. We have not run	799
764	existing detectors (DetectGPT, GPTZero) on	799
765	the same corpus.	799
766	5. Domain specificity. Currently implemented	799
767	for five physical-register domains. Extension	799
768	to abstract registers requires additional word	799
769	sets.	799
770	6. Passage length. Orphanhood tests operate on	799
771	passage-length windows. Earned figurative	799
772	language can score as orphaned if excerpted	799
773	from longer works.	799
774	7. Same-model variance. The Sonnet replica-	799
775	tion (15%) was individually non-significant	799
776	($p = 0.224$). Post-hoc power analysis shows	799
777	80% power to detect $h \geq 0.53$ at $n = 20$; the	799
778	replication’s $h = 0.39$ falls below this thresh-	799
779	old. The non-significant result reflects power	799
780	limitations, not absent signal.	799

800 was performed by deterministic rule-based algo-
 801 rithms.
 802

803 Code and Data Availability

804 Detection algorithms, experiment scripts, and cor-
 805 pus data will be released upon publication.
 806

807 References

- 808 Pat Barker. 1991. *Regeneration*. Viking.
 809
 810 Roland Barthes. 1970. *S/Z*. Éditions du Seuil.
 811
 812 Roland Barthes. 1973. *Le Plaisir du texte*. Éditions
 813 du Seuil.
 814
 815 Wendell Berry. 2000. *Jayber Crow*. Counterpoint.
 816
 817 Anthony Bourdain. 2000. *Kitchen Confidential*.
 818 Bloomsbury.
 819
 820 Bill Buford. 2006. *Heat*. Knopf.
 821
 822 Joseph Conrad. 1902. *Typhoon*. Heinemann.
 823
 824 Ruiqi Du, Maximilian Fung, Yifei Hu, and David
 825 Liu. 2026. Overcoming topology bias and cold-
 826 start limitations in drug repurposing: A clinical
 827 outcome-aligned LLM framework. *bioRxiv*.
 828
 829 M.F.K. Fisher. 1954. *The Art of Eating*. World
 830 Publishing.
 831
 832 Ernest Hemingway. 1952. *The Old Man and the
 833 Sea*. Scribner.
 834
 835 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes,
 836 and Yejin Choi. 2020. The curious case of neural
 837 text degeneration. In *Proceedings of ICLR 2020*.
 838
 839 Richard Hooker. 1968. *MASH: A Novel About
 840 Three Army Doctors*. Morrow.
 841
 842 Jaap Jumelet, Willem Zuidema, and Arabella Sin-
 843 clair. 2024. Syntactic structural priming in large
 844 language models. In *Proceedings of ACL 2024*.
 845
 846 Sebastian Junger. 1997. *The Perfect Storm*. Nor-
 847 ton.
 848
 849 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*.

- 850 Kalpesh Krishna, Yixiao Song, Marzena Karpinska,
 851 John Wieting, and Mohit Iyyer. 2023. Paraphras-
 852 ing evades detectors of AI-generated text, but
 853 retrieval is an effective defense. In *Proceedings
 854 of NeurIPS 2023*.
 855
 856 R. Kugler. 2025. Polysemy patterns in large lan-
 857 guage model output. *arXiv*. ArXiv:2511.21334.
 858
 Cormac McCarthy. 1985. *Blood Meridian*. Ran-
 859 dom House.
 860
 John McPhee. 1975. *The Survival of the Bark
 861 Canoe*. Farrar, Straus and Giroux.
 862
 Eric Mitchell, Yoonho Lee, Alexander Khazatsky,
 863 Christopher D. Manning, and Chelsea Finn.
 864 2023. DetectGPT: Zero-shot machine-generated
 865 text detection using probability curvature. In
 866 *Proceedings of ICML 2023*.
 867
 868 Masahiro Mori. 1970. The uncanny valley. *Energy*,
 869 7(4):33–35. K. F. MacDorman & N. Kageki,
 870 Trans., IEEE Robotics & Automation Magazine,
 871 19(2), 2012.
 872
 Patrick O’Brian. 1969. *Master and Commander*.
 873 Collins.
 874
 George Orwell. 1933. *Down and Out in Paris and
 875 London*. Gollancz.
 876
 Michael Pollan. 1997. *A Place of My Own*. Ran-
 877 dom House.
 878
 Kevin Powers. 2012. *The Yellow Birds*. Little,
 879 Brown.
 880
 Annie Proulx. 2016. *Barkskins*. Scribner.
 881
 882 Erich Maria Remarque. 1929. *Im Westen nichts
 883 Neues [All Quiet on the Western Front]*. Propy-
 884 läen.
 885
 Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-
 886 Olano, and Byron C. Wallace. 2025. Detection
 887 and measurement of AI-generated text quality
 888 dimensions: Expert taxonomy and span-level
 889 annotation. *arXiv*. ArXiv:2509.19163.
 890
 Dagmar Stumpfe and Jürgen Bajorath. 2012. Ex-
 891 ploring activity cliffs in medicinal chemistry.
 892 *Journal of Medicinal Chemistry*, 55(7):2932–
 893 2942.
 894
 George Sturt. 1923. *The Wheelwright’s Shop*.
 895 Cambridge University Press.
 896
 897

900	Flora Thompson. 1945. <i>Lark Rise to Candleford.</i>	950
901	Oxford University Press.	951
902		952
903	Edward Tian and 1 others. 2023. GPTZero: To-	953
904	wards detection of AI-generated text using zero-	954
905	shot and supervised methods. Preprint.	955
906		956
907	Ashish Vaswani, Noam Shazeer, Niki Parmar,	957
908	Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,	958
909	Łukasz Kaiser, and Illia Polosukhin. 2017. At-	959
910	tention is all you need. In <i>Proceedings of NeurIPS 2017.</i>	960
911		961
912	Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan,	962
913	Lidia S. Chao, and Derek F. Wong. 2025. A sur-	963
914	vey on LLM-generated text detection: Necessity,	964
915	methods, and future directions. <i>Computational</i>	965
916	<i>Linguistics</i> , 51(1):275–338.	966
917		967
918		968
919		969
920		970
921		971
922		972
923		973
924		974
925		975
926		976
927		977
928		978
929		979
930		980
931		981
932		982
933		983
934		984
935		985
936		986
937		987
938		988
939		989
940		990
941		991
942		992
943		993
944		994
945		995
946		996
947		997
948		998
949		999