

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

Anonymous ACL submission

001

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)

045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084

085 a semiotic interpretation connecting the signal to
086 the distinction between Barthes’s *significance* and
087 *signification*.

088 2 Related Work

089 2.1 AI Text Detection

090 Current detection methods include perplexity-
091 based classifiers (Mitchell et al., 2023), watermarking
092 (Kirchenbauer et al., 2023), and supervised clas-
093 sifiers trained on LLM output distributions (Tian
094 et al., 2023). These methods achieve variable ac-
095 curacy and degrade across domains, paraphrasing
096 attacks, and model updates (Krishna et al., 2023;
097 Wu et al., 2025). Critically, none provides a struc-
098 tural explanation for *what* distinguishes AI prose
099 from human prose at the level of craft.

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

104 2.2 Polysemy and Priming in LLM Output

105 Kugler (2025) demonstrates that LLM output ex-
106 hibits a “flatter semantic space” than natural lan-
107 guage (frequency-specificity correlation: $\rho \approx -0.3$
108 for LLMs vs. $\rho \approx -0.5$ to -0.7 for human text).
109 This flat distribution is consistent with our over-
110 indexing hypothesis. Jumelet et al. (2024) demon-
111 strate that lexico-semantic overlap boosts token-
112 level probability in transformers through structural
113 priming effects, confirming the mechanistic foun-
114 dation of our claim.

115 2.3 Coherence as a Latent Dimension

116 Shaib et al. (2025) develop a taxonomy of “AI slop”
117 through expert annotation, finding that standard
118 text metrics fail to capture coherence dimensions
119 and that capable LLMs likewise fail to reliably
120 identify slop. Our orphanhood framework provides
121 one structural answer: it operationalizes a specific
122 form of incoherence (figurative constructions arriv-
123 ing without architectural support) as a measurable,
124 deterministic signal.

125 3 Theoretical Framework

126 3.1 Latent Semantic Recruitment

127 We define *Latent Semantic Recruitment* (LSR) as
128 the phenomenon whereby an autoregressive lan-
129 guage model, generating text within an active fig-
130 urative register R , disproportionately selects pol-

131 ysemous words whose secondary semantic fields
132 overlap with R .

133 Let w be a word token with primary sense s_1
134 (contextually appropriate) and secondary sense s_2
135 (not contextually required). LSR occurs when
136 $P(w | \text{context}, R) > P(w | \text{context}, \neg R)$ specifi-
137 cally because the embedding of s_2 overlaps with R
138 in the model’s representation space.

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

150 3.2 Training-Weight Over-Indexing

151 LSR explains the mechanism, but not why the re-
152 sult is detectable. Human writers also select poly-
153 semous words. The critical question is why LLM
154 polysemous usage is distinguishable.

155 We propose the **training-weight over-indexing**
156 **hypothesis**: training corpora contain a distribu-
157 tional bias that systematically over-represents ex-
158 ceptional figurative prose. The texts exhibiting the
159 most sophisticated polysemous craft (Conrad, Mc-
160 McCarthy, Woolf, Morrison) receive the most analyti-
161 cal attention, pedagogical citation, and anthology
162 inclusion, producing the most duplication across
163 training data. Under standard cross-entropy train-
164 ing, tokens appearing more frequently contribute
165 proportionally more to the cumulative gradient.
166 The model therefore learns to reproduce this level
167 of sophistication not as exceptional but as the ex-
168 pected register of competent prose.

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

175 **Caveat.** A direct demonstration would require
176 measuring the frequency of specific figurative con-
177 structions in training data and correlating that fre-
178 quency with generation probability, an analysis re-
179 quiring training-data access we do not have. The
180 hypothesis is argued from distributional logic, con-
181 sistent with the observed signal, but not indepen-

182 dently verified.

183 3.3 Orphaned Sophistication

184 The over-indexing hypothesis predicts that LLMs
185 will produce sophisticated figurative language *with-
186 out the structural architecture that earns it*. We
187 define **orphaned sophistication** as a figurative
188 construction satisfying three conditions:

189 **Isolation.** The figurative density of the sen-
190 tence containing w is significantly higher than
191 its neighbors (window ± 2 sentences). Score:
192 $\min(1.0, (\varphi(s_w) - \bar{\varphi}(N(s_w)))/\tau_1)$, where $\tau_1 =$
193 0.2.

194 **Chain disconnection.** The register field acti-
195 vated by s_2 is not activated by other words within
196 ± 3 sentences. In human literary prose, figurative
197 constructions participate in metaphor chains. Score:
198 0 connections = 1.0, 1 = 0.6, 2 = 0.2, 3+ = 0.0.

199 **Lack of preparation.** The context is scored for
200 signposting markers: simile constructions, explicit
201 frame-setting, tonal shifts (sentence-length ratio >
202 2.5:1), and figurative density in adjacent sentences
203 (> 0.15). Score ranges from 1.0 (fully unprepared)
204 to 0.0 (fully prepared).

205 A word’s orphanhood score is the arithmetic
206 mean of its three test scores. A word exceeding
207 0.6 is classified as *orphaned*; words below that
208 threshold are classified as *structurally integrated*
209 (the figurative construction participates in the sur-
210 rounding prose architecture). The threshold was
211 set a priori, not optimized on test data.

212 3.4 Semiotic Interpretation

213 The framework admits an interpretation through
214 Barthes’s semiotic theory (Barthes, 1970, 1973),
215 though we acknowledge this narrows Barthes’s
216 framework considerably. In *S/Z*, *significance* con-
217 cerns the plurality of meaning generated by the
218 interaction of multiple codes in writerly texts; our
219 usage maps a more architectural reading onto the
220 term, treating *significance* as requiring structural
221 scaffolding. This narrowing is deliberate and oper-
222 ational.

223 When Conrad writes “the rudder would bite,” the
224 word performs something closer to *signification*: it
225 participates in a novel-length architecture. When
226 an LLM writes “the hungry steel teeth,” the same
227 semantic content is present but the structural labor
228 is absent. The word performs *signification* without
229 *significance*. Our three tests map onto this distinc-
230 tion: isolation measures sustained vs. anomalous

231 sophistication; chain connectivity measures pro-
232 ductive labor vs. standalone activation; preparation
233 measures deliberate register transition vs. its ab-
234 sence.

235 A necessary caveat: a sufficiently long LLM text
236 may, through stochastic density alone, produce pas-
237 sages scoring well on all three dimensions. Our
238 detector measures necessary conditions for *signifi-
239 ance* (structure is present) but not the sufficient
240 condition (structure was produced through autho-
241 rial labor). This is why we describe the framework
242 as identifying *orphaned* sophistication rather than
243 *unearned* sophistication.

244 4 Experimental Method

245 4.1 Corpus Construction

246 We assembled three corpora across five
247 physical-register domains (ocean storm,
248 kitchen/restaurant, blacksmith/forge, battle-
249 field surgery, sawmill/logging):

250 **Human Corpus A (Published).** 20 passages
251 (~100–200 words), drawn from published fiction
252 and nonfiction spanning 1902–2016, four per do-
253 main. Authors include Conrad (1902), Hemingway
254 (1952), McCarthy (1985), Bourdain (2000), Proulx
255 (2016), Powers (2012), Barker (1991), Remarque
256 (1929), Orwell (1933), and ten others.¹

257 **Human Corpus B (Non-professional).** 5 pas-
258 sages (~150–250 words), hand-written by a non-
259 professional writer under experimental conditions
260 (one per domain, written under time pressure with-
261 out revision, before the detection framework was
262 developed).

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

272 4.2 Corpus Provenance

273 Corpus A passages were reproduced from LLM
274 training data, introducing a potential circularity
275 (Section 7.5). Corpus B provides an uncontami-

276 ¹Full corpus: O’Brian (1969), Junger (1997), Buford
277 (2006), Fisher (1954), Thompson (1945), McPhee (1975),
278 Sturt (1923), Hooker (1968), Kesey (1964), Berry (2000),
279 Pollan (1997).

nated baseline. The detection instrument is deterministic (no LLM judgment in scoring), so circularity applies only to corpus construction.

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

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%

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

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.

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 $18.8 \times$ (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:

²Full 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	<i>n</i>	Rate	<i>h</i>	<i>p</i>
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%.

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 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.

Domain	Human	LLM	Fisher <i>p</i>
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	<i>h</i>	<i>p</i>
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.

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 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-

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	∞^*
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.

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 $1.91\times$).

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

491 specific words, explains why they are orphaned,
492 and provides structural explanation. For writers us-
493 ing LLMs collaboratively, the framework provides
494 actionable revision guidance: flagged passages re-
495 quire not deletion but *architecture* (build a chain,
496 prepare the register shift, sustain figurative den-
497 sity).

498 7.5 The Generalizable Principle

499 The hypothesis predicts that any domain where
500 models are trained on corpora dominated by ex-
501 ceptional exemplars will exhibit an analogous ar-
502 tifact. Du et al. (2026) independently identify
503 “hard negatives” in computational drug repurpos-
504 ing: well-studied compounds appearing ideal due
505 to high knowledge-graph connectivity but failing
506 clinically. The mechanism is structurally identical:
507 FDA-approved drugs dominate training corpora
508 through citation and patent literature, producing
509 the same over-indexing dynamic. Graph neural
510 networks learn to produce binding moieties resem-
511 bling successful drugs as default output, locally
512 brilliant binding predictions structurally orphaned
513 from the ADMET properties that would make them
514 clinically viable. This corresponds to the “activ-
515 ity cliff” problem in medicinal chemistry (Stumpfe
516 and Bajorath, 2012).

517 The mapping is exact: a high-affinity binding
518 moiety without metabolic stability is an orphaned
519 figurative word. Isolation, chain disconnection, and
520 lack of preparation all have molecular analogues.
521 The convergence, observed independently in a dif-
522 ferent modality by researchers with no knowledge
523 of the orphanhood framework, suggests the pathol-
524 ogy is a general property of training-data distribu-
525 tional bias. (We note that Du et al. is a bioRxiv
526 preprint not yet peer-reviewed.)

527 Limitations

- 528 1. **Sample size.** $n = 125$ (25 human, 100 LLM)
529 across 5 domains and 3 families. Statistically
530 significant but modest for strong generalization.
- 532 2. **Human corpus provenance.** Corpus A was
533 reproduced from LLM training data. Cor-
534 pus B provides an uncontaminated baseline,
535 but ideal replication would use passages trans-
536cribed from physical books.
- 537 3. **Human corpus skill ceiling.** The detector’s
538 FPR was measured against elite literary prose.

Its behavior on mid-tier prose (workshop fic-
539 tion, genre fiction, journalism) is unknown
540 and is a priority for future validation.

- 542 4. **No baseline comparison.** We have not run
543 existing detectors (DetectGPT, GPTZero) on
544 the same corpus.
- 545 5. **Domain specificity.** Currently implemented
546 for five physical-register domains. Extension
547 to abstract registers requires additional word
548 sets.
- 549 6. **Passage length.** Orphanhood tests operate on
550 passage-length windows. Earned figurative
551 language can score as orphaned if excerpted
552 from longer works.
- 553 7. **Same-model variance.** The Sonnet replica-
554 tion (15%) was individually non-significant
555 ($p = 0.224$). Post-hoc power analysis shows
556 80% power to detect $h \geq 0.53$ at $n = 20$; the
557 replication’s $h = 0.39$ falls below this thresh-
558 old. The non-significant result reflects power
559 limitations, not absent signal.

560 8 Conclusion

561 We have identified and formalized a novel artifact
562 of autoregressive language generation: orphaned
563 sophistication. In pooled analysis across three in-
564 dependent model families, our experiments pro-
565 vide statistically significant evidence for the artifact
566 (Fisher’s $p = 0.006$, Cohen’s $h = 0.71$, $n = 125$).
567 The artifact arises, we argue, from training-weight
568 over-indexing on exceptional exemplars, causing
569 models to produce locally sophisticated outputs
570 without the structural integrity that would earn
571 them.

572 The uncanny valley of AI prose is a measurable
573 structural coherence failure. The detector identi-
574 fies where figurative constructions lack architec-
575 tural support and what structural work would close
576 the gap. Whether the signal persists under fine-
577 tuning, extends to abstract registers, or inverts in
578 constrained-vocabulary environments where com-
579 pilers enforce integration remains to be tested.

580 LLM Assistance Disclosure

581 This paper was written with the assistance of sev-
582 eral large language models used as research tools.
583 Claude Opus 4 (Anthropic) was used to implement
584 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

- 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.
- 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.