

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

Anonymous ACL submission

001

Abstract

002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044

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

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

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)

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. Classification thresh-
207 old: 0.6 (set a priori, not optimized on test data).

208 3.4 Semiotic Interpretation

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

219 When Conrad writes “the rudder would bite,” the
220 word performs something closer to *significance*: it
221 participates in a novel-length architecture. When
222 an LLM writes “the hungry steel teeth,” the same
223 semantic content is present but the structural labor
224 is absent. The word performs *signification* without
225 *significance*. Our three tests map onto this distinc-
226 tion: isolation measures sustained vs. anomalous
227 sophistication; chain connectivity measures pro-
228 ductive labor vs. standalone activation; preparation
229 measures deliberate register transition vs. its ab-
230 sence.

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

240 4 Experimental Method

241 4.1 Corpus Construction

242 We assembled three corpora across five
243 physical-register domains (ocean storm,
244 kitchen/restaurant, blacksmith/forge, battle-
245 field surgery, sawmill/logging):

246 **Human Corpus A (Published).** 20 passages
247 (~ 100 –200 words), drawn from published fiction
248 and nonfiction spanning 1902–2016, four per do-
249 main. Authors include Conrad (1902), Hemingway
250 (1952), McCarthy (1985), Bourdain (2000), Proulx
251 (2016), Powers (2012), Barker (1991), Remarque
252 (1929), Orwell (1933), and ten others.¹

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

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

268 4.2 Corpus Provenance

269 Corpus A passages were reproduced from LLM
270 training data, introducing a potential circularity
271 (Section 7.5). Corpus B provides an uncontami-
272 nated baseline. The detection instrument is deter-
273 ministic (no LLM judgment in scoring), so circu-
274 larity applies only to corpus construction.

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).

275 4.3 Detection Instrument

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

286 The reported instrument (detector v3) identifies
 287 figurative polysemous words using v2’s mech-
 288 anisms, then subjects each candidate to the three
 289 orphanhood tests defined in Section 3.3. The algo-
 290 rithm processes each sentence, identifying words
 291 that (a) are not in the domain-literal set for the
 292 passage’s domain (34–41 words per domain), (b)
 293 appear in at least one of six register fields (con-
 294 sumption, personification, body, violence, fire/heat,
 295 water/weather; 16–29 words each), and (c) exhibit
 296 figurative usage (personification, animate verb, or
 297 animate-quality modifier). Qualifying words re-
 298 ceive the three orphanhood scores; the arithmetic
 299 mean must exceed 0.6 for classification as or-
 300 phaned.

301 The detector is fully deterministic: no neural
 302 network, no LLM, no learned parameters. All
 303 thresholds are set a priori. Domain-literal filter-
 304 ing is conservative by design: it suppresses only
 305 words whose primary sense denotes the domain
 306 activity. A word like “grip” is not added to the
 307 blacksmith domain-literal set even though black-
 308 smiths literally grip tools, because its primary sense
 309 (physical grasping) is not specific to blacksmithing
 310 (see supplementary material E.4). Domain-literal
 311 filtering operates on the passage’s declared domain,
 312 assigned at corpus construction time.²

313 4.4 Statistical Methods

314 All comparisons use Fisher’s exact test (appropri-
 315 ate for small-sample count data). We report one-
 316 sided p -values (testing the directional hypothesis
 317 that LLM rates exceed human rates) and two-sided
 318 p -values. Confidence intervals use the Clopper-
 319 Pearson exact method ($\alpha = 0.05$). Effect sizes are
 320 Cohen’s h , where $h > 0.8$ is conventionally large.
 321 No multiple-comparison correction is applied to
 322 the primary analysis (single pre-specified compari-

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 $h = 0.86$.

son); per-domain exploratory analyses are flagged
 323 as uncorrected.

325 4.5 Experimental Design

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

337 5 Results

338 5.1 Primary Analysis

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

343 The detector v2 analysis (unjustified figurative
 344 polysemy, before orphanhood filtering) showed a
 345 rate ratio of $18.8 \times$ (LLM 0.750 per passage vs. hu-
 346 man 0.040), Fisher’s $p = 0.001$, Cohen’s $h = 1.69$.
 347 The v3 orphanhood model correctly reclassified 6
 348 of the 15 v2 detections as “earned”: cases where the
 349 model had accidentally produced chain connectiv-
 350 ity (e.g., “bit” appearing near “teeth” or “hungry”).

351 5.2 Qualitative Analysis

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

355 By contrast, human passages employing figu-
 356 rative language do so within explicitly prepared
 357 frames. The non-professional sawmill passage
 358 (“Life shaves pieces of your health off... Bertha
 359 takes that, too”) signposts the saw-as-life metaphor
 360 (“That’s life, that is”), develops it across multiple

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%.

361 sentences, and connects to a chain of related vocabulary. The detector correctly identifies this as
362 earned.
363

364 5.3 Cross-Model and Cross-Family Validation

365 Three of five model runs are individually significant.
366 The pooled analysis across three independent
367 families provides the definitive test ($p = 0.006$,
368 $h = 0.71$, power 93.8%).

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

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

382 5.4 Per-Domain Distribution

383 The kitchen domain produced a null result (0/5
384 human, 0/4 LLM), the only domain with zero de-
385 tections. Culinary language is inherently action-
386 oriented and consumption-related, so words that
387 would register as figurative in other domains are
388 domain-literal in a kitchen context. The detector

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.

389 correctly identifies these as non-figurative.

390 5.5 Ablation Study

391 The results reveal an asymmetric architecture. Re-
392 moving chain connectivity reduces *h* by 0.141, con-
393 firming that chain detection captures discriminative
394 signal. Removing preparation produces the largest
395 *h* degradation (0.193) and doubles the human false
396 positive rate from 4.0% to 8.0%, indicating that
397 preparation is the dimension most responsible for
398 specificity.

399 Removing isolation *increases* *h* to 0.821 because
400 isolation functions as a conservative filter, sup-
401 pressing true positives where the figurative spike
402 coincides with mildly elevated neighborhood den-
403 sity. All four configurations maintain significance
404 ($p < 0.05$).

405 6 Mechanism Validation

406 6.1 Monte Carlo Logit Proxy

407 A Monte Carlo simulation generated 100,000 ran-
408 dom word-context pairings and computed register-
409 overlap scores. Of 15 register-field/domain pair-
410 ings, 12 showed zero overlap between the random
411 distribution and the observed LLM orphan scores
412 ($p < 10^{-5}$ each), confirming that the observed
413 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	∞^*
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.

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 specific words, explains why they are orphaned, and provides structural explanation. For writers using LLMs collaboratively, the framework provides

488 actionable revision guidance: flagged passages re-
489 quire not deletion but *architecture* (build a chain,
490 prepare the register shift, sustain figurative den-
491 sity).

492 7.5 The Generalizable Principle

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

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

521 Limitations

- 522 1. **Sample size.** $n = 125$ (25 human, 100 LLM)
523 across 5 domains and 3 families. Statistically
524 significant but modest for strong generalization.
- 526 2. **Human corpus provenance.** Corpus A was
527 reproduced from LLM training data. Corpus B provides an uncontaminated baseline,
528 but ideal replication would use passages trans-
529 cribed from physical books.
- 531 3. **Human corpus skill ceiling.** The detector’s
532 FPR was measured against elite literary prose.
533 Its behavior on mid-tier prose (workshop fic-
534 tion, genre fiction, journalism) is unknown
535 and is a priority for future validation.

4. **No baseline comparison.** We have not run
536 existing detectors (DetectGPT, GPTZero) on
537 the same corpus.
538

5. **Domain specificity.** Currently implemented
539 for five physical-register domains. Extension
540 to abstract registers requires additional word
541 sets.
542

6. **Passage length.** Orphanhood tests operate on
543 passage-length windows. Earned figurative
544 language can score as orphaned if excerpted
545 from longer works.
546

7. **Same-model variance.** The Sonnet replica-
547 tion (15%) was individually non-significant
548 ($p = 0.224$). Post-hoc power analysis shows
549 80% power to detect $h \geq 0.53$ at $n = 20$; the
550 replication’s $h = 0.39$ falls below this thresh-
551 old. The non-significant result reflects power
552 limitations, not absent signal.
553

554 8 Conclusion

We have identified and formalized a novel artifact
555 of autoregressive language generation: orphaned
556 sophistication. Our experiments provide statisti-
557 cally significant evidence across three indepen-
558 dent model families (Fisher’s $p = 0.006$, Cohen’s
559 $h = 0.71$, $n = 125$). The artifact arises, we argue,
560 from training-weight over-indexing on exceptional
561 exemplars, causing models to produce locally so-
562 phisticated outputs without the structural architec-
563 ture that would earn them.
564

The detection signal is structural and inter-
565 pretable: it provides not merely a classification
566 but a diagnosis of where the architecture is missing
567 and what kind of work would repair it. The core
568 contribution is a reframing: the uncanny valley of
569 AI prose is a structural coherence failure, not a
570 lexical quality failure, and it is measurable. The
571 machine does not write badly. It writes too well, in
572 moments that have not been earned.
573

574 LLM Assistance Disclosure

This paper was written with the assistance of sev-
575 eral large language models used as research tools.
576 Claude Opus 4 (Anthropic) was used to implement
577 detection algorithms, reproduce published passages
578 from training data for Corpus A, generate statisti-
579 cal analyses, and draft the manuscript under the
580 author’s direction. Claude Sonnet 4, Claude Haiku
581 3.5, GPT-4o, and Gemini 2.5 Flash generated the
582

583 respective LLM test corpora. No model involved in
584 corpus generation was involved in scoring: all de-
585 tection was performed by deterministic rule-based
586 algorithms.

587 References

- 588 Pat Barker. 1991. *Regeneration*. Viking.
589 Roland Barthes. 1970. *S/Z*. Éditions du Seuil.
590 Roland Barthes. 1973. *Le Plaisir du texte*. Éditions du
591 Seuil.
592 Wendell Berry. 2000. *Jayber Crow*. Counterpoint.
593 Anthony Bourdain. 2000. *Kitchen Confidential*.
594 Bloomsbury.
595 Bill Buford. 2006. *Heat*. Knopf.
596 Joseph Conrad. 1902. *Typhoon*. Heinemann.
597 Ruiqi Du, Maximilian Fung, Yifei Hu, and David Liu.
598 2026. Overcoming topology bias and cold-start lim-
599 itations in drug repurposing: A clinical-outcome-
600 aligned LLM framework. *bioRxiv*.
601 M.F.K. Fisher. 1954. *The Art of Eating*. World Publish-
602 ing.
603 Ernest Hemingway. 1952. *The Old Man and the Sea*.
604 Scribner.
605 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and
606 Yejin Choi. 2020. The curious case of neural text
607 degeneration. In *Proceedings of ICLR 2020*.
608 Richard Hooker. 1968. *MASH: A Novel About Three
609 Army Doctors*. Morrow.
610 Jaap Jumelet, Willem Zuidema, and Arabella Sinclair.
611 2024. Syntactic structural priming in large language
612 models. In *Proceedings of ACL 2024*.
613 Sebastian Junger. 1997. *The Perfect Storm*. Norton.
614 Ken Kesey. 1964. *Sometimes a Great Notion*. Viking.
615 John Kirchenbauer, Jonas Geiping, Yuxin Wen,
616 Jonathan Katz, Ian Miers, and Tom Goldstein. 2023.
617 A watermark for large language models. In *Proceed-
618 ings of ICML 2023*.
619 Kalpesh Krishna, Yixiao Song, Marzena Karpinska,
620 John Wieting, and Mohit Iyyer. 2023. Paraphrasing
621 evades detectors of AI-generated text, but retrieval
622 is an effective defense. In *Proceedings of NeurIPS
623 2023*.
624 R. Kugler. 2025. Polysemy patterns in large language
625 model output. *arXiv*. ArXiv:2511.21334.
626 Cormac McCarthy. 1985. *Blood Meridian*. Random
627 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 detec-
tion 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 mea-
surement 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*. Cam-
bridge University Press.
Flora Thompson. 1945. *Lark Rise to Candleford*. Ox-
ford 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, meth-
ods, and future directions. *Computational Linguis-
tics*, 51(1):275–338.