

# Depth Profiling: Classifying Recursive Generation Depth in Base-Model LLM Text

Apex Intelligence

## Abstract

As large language model (LLM) outputs re-enter training corpora and seed further generation, understanding how recursive generation reshapes text distributions is critical for data provenance and corpus curation. Existing detectors answer only the binary human-or-machine question, ignoring *generation depth*; we introduce *depth profiling* to address this gap. We decompose distributional drift under recursive generation into location and shape components. Across eight base-model generators from five families, negative log-likelihood (NLL) location features drift monotonically with depth while shape features remain near-stationary—an effect-size gap exceeding  $8\times$  ( $d=0.74\text{--}1.05$  vs.  $d < 0.10$ ). We term this asymmetry the *differential depth response* (DDR). Prior work documents *what converges*; the central finding here is *what is approximately preserved*: shape stability complements location drift for ordinal depth grading without per-generator calibration. A lightweight ensemble achieves leak-free binary  $F1=0.694$  under 8-generator leave-one-generator-out evaluation; cross-model chains reach  $F1=0.830$ . Instruction tuning compresses DDR and bounds current applicability to base models<sup>1</sup>.

## 1 Introduction

An estimated 7–17% of recent scientific publications already contain substantial LLM-modified content (Liang et al., 2024b,a). As these outputs re-enter training corpora (Soldaini et al., 2024; Penedo et al., 2024) and seed further generation, text accumulates at varying *generation depths*; each successive cycle reshapes distributional properties in ways that matter for data provenance, corpus auditing, and contamination monitoring in base-model pretraining corpora (Wenzek et al., 2020; Xie et al.,

2023) as well as model collapse analysis (Shumailov et al., 2024; Dohmatob et al., 2025). Understanding *how* recursive generation alters text distributions, and which properties change versus which are preserved, is therefore a prerequisite for principled corpus curation, yet current methods address only binary human-or-machine detection (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024; Narayanasamy et al., 2026). Base models provide the cleanest substrate to isolate depth effects from alignment-induced changes. We find that recursive generation produces a decomposable depth profile: distributional *location* drifts with depth while *shape* remains stable across generators, an asymmetry sufficient for cross-generator depth prediction.

Distributional compression under recursive generation is well-documented (Guo et al., 2024; Mohamed et al., 2025; Geng et al., 2026), yet existing analyses characterize *that* and *why* convergence occurs without decomposing what drifts from what is preserved. We introduce *depth profiling* (coarse ordinal classification into gen-0/gen-1/gen-2+) and decompose this compression into structurally distinct components. Across eight base-model generators spanning five architectural families (0.6B–8B), the NLL distribution’s central tendency—mean entropy (per-token NLL under a fixed reference model, not information-theoretic entropy; §3.3) and lexical diversity—drifts monotonically with depth, while its shape (entropy extremes, tail structure) remains  $\geq 8\times$  more stable ( $d=0.74\text{--}1.05$  vs.  $d < 0.10$ ; entropy std is intermediate at  $d=0.46$ ). We term this location/shape asymmetry the *differential depth response* (DDR; Figure 1), an empirical regularity of base-model autoregressive generation that is mechanistically distinct from model collapse (Shumailov et al., 2024). Model collapse erodes tail coverage via weight updates; DDR instead analyzes fixed-parameter inference chains. The stationary shape component complements location drift for depth

<sup>1</sup>Code and data are available at <https://anonymous.4open.science/r/gen-depth-fingerprint-3FB3/>.

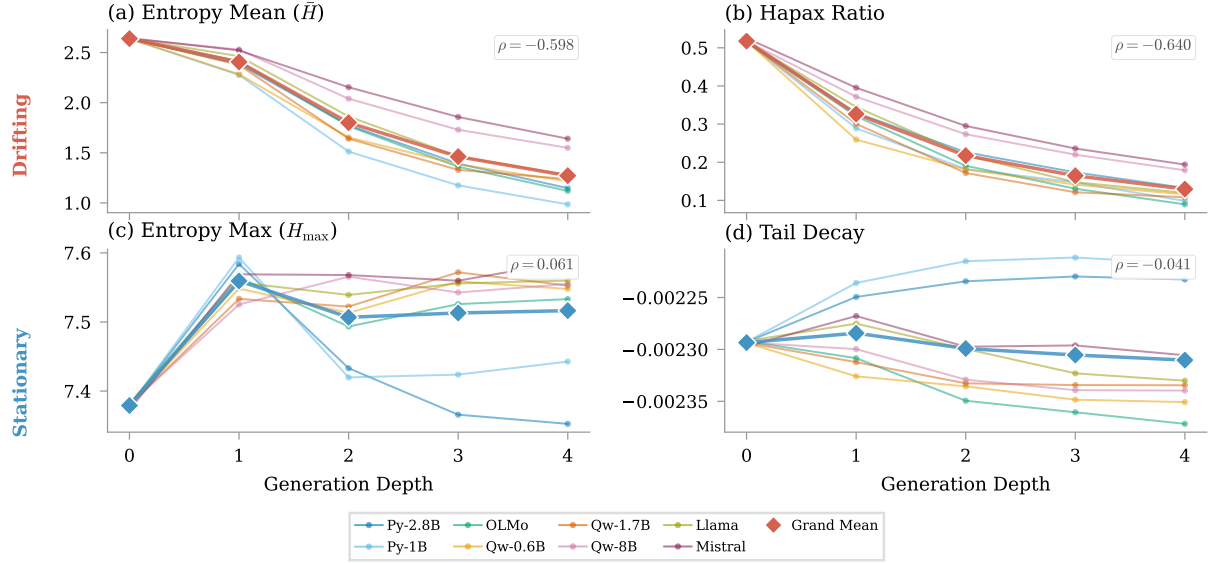


Figure 1: Feature trajectories across generation depth ( $y$ -axis:  $z$ -scored feature values). **Top row (drifting)**: entropy mean and hapax ratio decline monotonically across all eight generators ( $|\rho| > 0.59$ ). **Bottom row (stationary)**: entropy max and tail decay remain flat ( $|\rho| < 0.07$ ). Colored lines: individual generators; diamonds: cross-generator mean (red for drifting, blue for stationary). This dichotomy defines the DDR.

prediction without per-generator calibration.

We focus on base (pretrained) models; instruction tuning degrades depth signal (Appendix k). Figure 2 summarizes the pipeline: recursive generation chains (§3.2), feature extraction with DDR partitioning (§3.3), and leave-one-generator-out (LOGO) evaluation (§3.4).

As a proof of concept, a lightweight 7-feature ensemble informed by DDR achieves leak-free binary F1 = 0.694 (macro-averaged LOGO; pooled: 0.705), matching multi-detector ensembles (0.708) at  $\sim 2.4\times$  lower cost; an end-to-end QLoRA (Dettmers et al., 2023) (Quantized Low-Rank Adaptation) classifier reaches 3-class F1 = 0.905 (within-model 5-fold CV; Appendix e), and the depth signal thus extends well beyond handcrafted features.

Our contributions:

1. **Empirical decomposition (DDR)**. Across eight generators from five families, we decompose inference-chain drift into two primary response zones (drifting,  $d=0.74$ – $1.05$ ; near-stationary,  $d < 0.10$ ) plus one intermediate feature ( $d=0.46$ ). Where prior work documents *what converges* (Guo et al., 2024; Geng et al., 2026), DDR identifies *what is approximately preserved*, an empirical regularity with analytical utility for depth classification without per-generator calibration.
2. **Task formulation and evaluation suite**. We

define generation depth classification as an ordinal prediction task and construct a controlled dataset of 37K texts from eight generators across five families (0.6B–8B; exploratory evidence extends to 14B, Appendix z.7), with leave-one-generator-out cross-validation as the primary evaluation protocol.

3. **Scope characterization**. Systematic evaluation across temperature, prompt type, instruction tuning, model scale, domain transfer, and perturbation robustness maps the operational envelope of depth profiling.

## 2 Related Work

**AI-generated text detection.** Detection methods span zero-shot scoring (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024; Yang et al., 2024), learned classifiers (Verma et al., 2024; Narayanasamy et al., 2026; Sadasivan et al., 2025), and surprisal-based approaches (Basani and Chen, 2026; Xu et al., 2024); see Wu et al. (2025) for a survey. Recent work moves beyond binary detection—multi-generator attribution (Wang et al., 2024), collaboration roles (Cheng et al., 2025), joint detection-attribution (Rao et al., 2025)—but all taxonomize by *source* or *production mode*, not recursive generation *depth*. Recent work reframes detection as an out-of-distribution problem (Zeng et al., 2025); depth profiling differs in targeting ordinal depth rather than binary membership, with

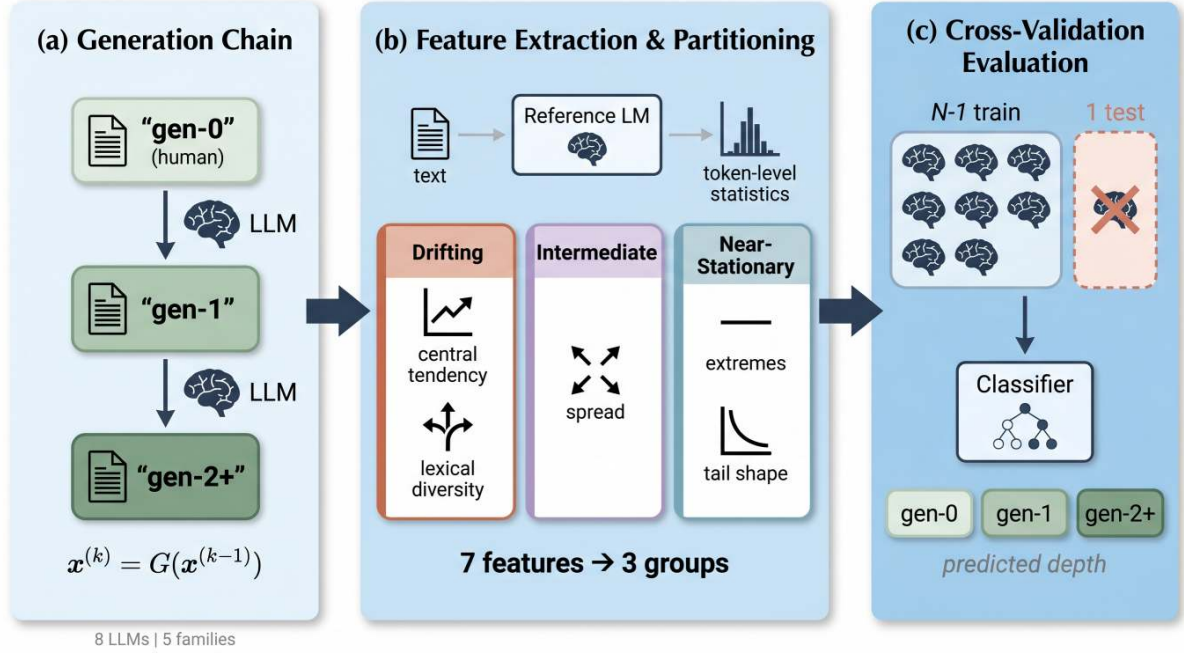


Figure 2: Depth profiling pipeline. (a) Human seeds are iteratively fed through generator  $G$  to produce chains of increasing depth. (b) Seven features extracted from NLL distributions group into three DDR zones: drifting (red), intermediate (purple), near-stationary (blue). (c) LOGO evaluation holds out one generator for testing.

cross-generator transfer as the evaluation axis. We introduce ordinal depth classification (Gutiérrez et al., 2016; Cao et al., 2020) as a complementary task; repurposing existing detection scores as depth predictors reveals high cross-generator variance (binary F1 = 0.54–0.71; Table 3).

**Model collapse and iterative generation.** Self-consuming generative models lose distributional coverage via training-loop parameter updates (Alemohammad et al., 2024; Shumailov et al., 2024; Dohmatob et al., 2025, 2024; Schaeffer et al., 2025); we instead examine *inference chains* where parameters are fixed. Concurrent iterative-generation studies document complementary phenomena: paraphrasing chains show factuality degradation (Mohamed et al., 2025), Markovian analysis bounds distinguishable depth to  $\sim 3$  classes (Geng et al., 2026), and attractor cycle evidence emerges (Wang et al., 2025). DDR differs in chain type (non-paraphrastic continuation), analysis lens (location/shape decomposition identifying what is *preserved*), and downstream utility (shape stability enables cross-generator classification without calibration); a detailed comparison appears in Appendix z.8.

**Diversity, stylometry, and provenance.** Declining lexical diversity under recursive generation

(Guo et al., 2024; Padmakumar and He, 2024) aligns with the monotonic TTR and hapax ratio decline we observe; these classical stylometric features (Stamatatos, 2009) constitute DDR’s drifting component. Data curation relies on quality signals (Xie et al., 2023; Wenzek et al., 2020); generation depth offers a complementary provenance axis. Watermarking (Kirchenbauer et al., 2023; Dathathri et al., 2024) requires generator cooperation and degrades across multi-step chains (Ameen et al., 2026); our post-hoc approach requires no generation-time modification. Our NLL tail analysis extends probability-based membership inference (Shi et al., 2024; Koike et al., 2025) from binary membership to ordinal depth.

### 3 Method

#### 3.1 Task Formulation

Given a text  $x$ , we predict its *generation depth*  $d \in \{0, 1, 2+\}$ : human-written ( $d=0$ ), first-generation LLM output ( $d=1$ ), or text produced by two or more recursive LLM generations ( $d=2+$ ). A generation chain of depth  $k$  is defined recursively:  $x^{(0)} \sim \mathcal{C}_0$  (human corpus),  $x^{(k)} = G(x^{(k-1)})$  for  $k \geq 1$ , where  $G$  denotes stochastic autoregressive generation. Depths beyond 2 are merged into  $2+$  because feature convergence compresses inter-class margins at deeper levels (§4.2).

Table 1: Seven distributional features. “Entropy” features are per-token NLL under Pythia-410M (§3.3); surface features require no reference model. DDR analysis (§4.2) reveals entropy mean, TTR, and hapax ratio drift with depth while entropy max, min, and tail decay remain near-stationary; entropy std is intermediate.

| Feature            | Source   | Measures                            |
|--------------------|----------|-------------------------------------|
| entropy_mean (NLL) | Ref. NLL | Central tendency of token surprise  |
| entropy_std (NLL)  | Ref. NLL | Spread of token surprise            |
| entropy_max (NLL)  | Ref. NLL | Maximum token surprise              |
| entropy_min (NLL)  | Ref. NLL | Minimum token surprise              |
| tail_decay         | Ref. NLL | Log-prob decay slope (ranks 100–1K) |
| ttr                | Surface  | Type-token ratio                    |
| hapax_ratio        | Surface  | Proportion of hapax legomena        |

### 3.2 Generation Chain Construction

We construct generation chains using eight generators spanning five model families: **Pythia** (2.8B, 1B) (Biderman et al., 2023), **OLMo** (OLMo-2-1B) (Groeneveld et al., 2024), **Qwen** (Qwen3-1.7B, Qwen3-0.6B, Qwen3-8B-Base) (Qwen Team, 2025), **Llama** (Llama-3.2-3B) (Grattafiori et al., 2024), and **Mistral** (Mistral-7B-v0.3) (Jiang et al., 2023). These span a  $13\times$  parameter range (0.6B–8B) across five architectures; Qwen’s overrepresentation (3/8) is a limitation mitigated by leave-one-family-out (LOFO) evaluation ( $F1 = 0.686$ ; §4.1).

For each generator, we sample approximately 1,000 seed passages from English Wikipedia as  $C_0$ . The same seed pool serves as gen-0 for all generators, controlling for input content; in LOGO evaluation this creates a potential train–test overlap for gen-0, addressed by our binary evaluation excluding gen-0 entirely (§4.3). Each seed is iteratively fed through the generator to produce a chain of depth 4 (i.e., gen-0 through gen-4), using the prompt “Continue the following text: {text}”. We generate up to depth 4 to empirically confirm convergence: 5-class LOGO yields macro  $F1 \approx 0.50$ , with depths 2–4 near-indistinguishable (gen-2/gen-3 per-class  $F1 < 0.27$ ; Appendix d). This motivates the gen-2+ class collapse. All generation uses temperature  $\tau=0.7$ , nucleus sampling with  $p=0.9$ , and maximum length of 256 new tokens. After filtering chain breaks and passages too short for feature extraction, 37,058 texts remain across all depths (Appendix a); depths 2–4 are collapsed into 2+.

### 3.3 Feature Extraction

**Terminology.** Throughout this paper, we refer to per-token cross-entropy loss under a fixed reference model as “entropy” for conciseness; these are reference-model NLL values, not true information-theoretic entropy of the generating distribution.

We use interpretable distributional statistics so that the DDR partition remains transparent; end-to-end neural features improve classification (+10.0 pp; Appendix e) but obscure the decomposition (LOFO confirms both feature groups contribute independently; Appendix f).

We extract seven distributional features per text (Table 1), covering the NLL distribution’s location, spread, extremes, and tail structure, plus two reference-free surface diversity measures; six additional NLL statistics were rejected for redundancy ( $|r| > 0.65$ ,  $\Delta F1 \leq -0.011$ ; Appendix f.1).

**Reference-model features.** A held-out reference model (Pythia-410M, chosen to be smaller than all generators) computes per-token negative log-likelihood (NLL) for each text. Cross-family reference validation (Appendix s) shows  $F1$  spread of only 0.013 across Pythia, OLMo, and Qwen references, so DDR is reference-model agnostic; however, all tested references are causal LMs. From the resulting NLL sequence, we extract five statistics: mean, standard deviation, maximum, and minimum of the token-level NLL, plus tail\_decay (the slope of the vocabulary-level averaged log-probability distribution in ranks 100–1000), which captures the power-law regime where depth-induced compression is most detectable.

**Surface features.** Two features capture lexical diversity without requiring a reference model: type-token ratio (TTR) and the proportion of words appearing exactly once (hapax\_ratio).

**Feature complementarity.** These features exhibit a *differential depth response* (DDR), separating into drifting and non-drifting groups (§4.2; pairwise correlations in Appendix r). TTR and hapax ratio are strongly correlated ( $r = 0.97$ ); both are retained for marginal LOFO gains.

### 3.4 Evaluation Protocol

**Cross-validation.** We evaluate generalization via **LOGO** cross-validation: each fold holds out one generator entirely, so the classifier must generalize to a generator absent from training. This constitutes the strictest per-model protocol for assessing

deployment against previously unseen generators. LOGO does not isolate family-level effects; **LOFO** addresses this by holding out entire model families (binary F1 = 0.686; §4.1). Generator subsets vary across experiments (Appendix w).

**Classifier.** We use XGBoost (Chen and Guestrin, 2016) throughout ( $n_{\text{estimators}}=200$ ,  $\text{max\_depth}=6$ , learning rate 0.1; fixed without tuning), with identical hyperparameters across all methods in Table 3. Alternative classifiers (CORAL (Cao et al., 2020), CORN (Shi et al., 2023), Frank-Hall (Frank and Hall, 2001)) yield comparable performance (Appendix h); hyperparameter sensitivity ( $\text{max\_depth}$  sweep) shows F1 variation  $\leq 2.0$  pp, so discriminative power resides in the features, not the classifier.

**Metrics and statistical testing.** The primary metric is **macro-averaged F1** over the three depth classes, with 95% bootstrap confidence intervals (CIs;  $n=1,000$ ). All  $p$ -values are two-sided with Bonferroni correction within each family of simultaneous tests.

**Baselines.** We compare the 7-feature ensemble against four single-score detection methods repurposed as depth predictors: Avg-LogProb, FastDetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), and LD-Score (Narayanasamy et al., 2026). Each baseline score is used as a univariate feature in the same XGBoost framework under LOGO cross-validation to ensure a controlled comparison.

**Length confound control.** Since TTR and  $\text{hax\_ratio}$  are mechanically length-dependent, we verify depth signal is not a length artifact via length-matched subsampling that equalizes length distributions across depth classes (§4.1; Appendix b).

## 4 Results

### 4.1 Depth Classification Performance

The 7-feature ensemble achieves 3-class LOGO F1 = 0.762 (95% CI [0.758, 0.766];  $N=37,058$ ) across eight generators from five families (Table 2); cross-generator F1 spread of 9.1 pp captures generator selection uncertainty. This 3-class metric is inflated by shared gen-0 seeds (per-class gen-0 F1  $\approx 0.94$ ; §3.2); we therefore report **leak-free binary F1** (gen-1 vs. gen-2+) = 0.694 (CI [0.688, 0.700]) throughout, excluding gen-0 entirely. Per-generator binary F1 ranges from 0.633 (Mistral-7B) to 0.724 (Pythia-2.8B; std.

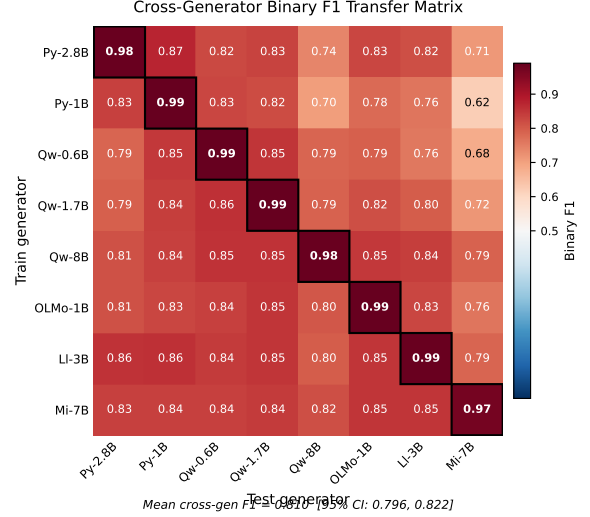


Figure 3: Cross-generator 8×8 transfer matrix (binary F1, gen-1 vs. gen-2+). Mean cross-generator F1 = 0.810 [95% CI: 0.796, 0.822]; uniformly high off-diagonal values confirm depth signal transfers across five families without calibration.

dev. 0.028), with the primary difficulty at the gen-1 / gen-2+ boundary (Appendix g).

**Cross-model generalization.** Cross-model chains (randomly selected generator per step) yield binary F1 = 0.830 (95% CI [.818, .841]; Appendix n) under 5-generator LOGO.<sup>2</sup> Single-generator transfer confirms the signal is not a mixed-chain artifact (mean transfer F1 = 0.810; Figure 3; full 8×8 matrix: Appendix y; LOFO = 0.686). ANOVA confirms generator identity contributes  $<1\%$  of feature variance ( $\eta_{\text{gen}}^2 \leq 0.009$ ; Appendix n). A scale effect emerges: smaller models achieve higher F1 (Pythia-2.8B: 0.724 vs. Mistral-7B: 0.633), partly due to same-family affinity with the Pythia-410M reference (spread  $\leq 0.013$ ; §4.2). An end-to-end QLoRA classifier reaches 3-class F1 = 0.905 (within-model 5-fold CV; +10.0 pp; Appendix e), evidence that additional depth signal resides in token-level representations. On instruct data, QLoRA recovers to 0.89–0.91 (within-model; Appendix k).

**Confound controls.** Length-matched subsampling yields binary F1 = 0.724 (vs. 0.770; Appendix b); independent-seed evaluation with disjoint pools (F1 = 0.680; Appendix t), so the classifier captures generative degradation rather than seed-specific content.

<sup>2</sup>Pythia-2.8B/1B, OLMo-2-1B, Qwen3-0.6B/1.7B; generator-level LOGO with bootstrap CIs ( $n=1,000$ ).

Table 2: 8-generator LOGO macro-F1 (bootstrap 95% CIs,  $n = 1,000$ ). Binary F1 excludes gen-0. <sup>†</sup>3-class F1 has upward bias from shared gen-0 seeds; independent-seed controls confirm the signal (F1=0.680; Appendix t).

| Model                       | Family  | Params | Bin F1       | 95% CI       | 3-class F1 <sup>†</sup> |
|-----------------------------|---------|--------|--------------|--------------|-------------------------|
| Pythia-2.8B                 | Pythia  | 2.8B   | <b>0.724</b> | [.707, .742] | 0.798                   |
| Pythia-1B                   | Pythia  | 1.0B   | 0.718        | [.700, .735] | <b>0.806</b>            |
| Qwen3-1.7B                  | Qwen    | 1.7B   | 0.709        | [.693, .726] | 0.772                   |
| Llama-3.2-3B                | Llama   | 3.2B   | 0.697        | [.680, .713] | 0.775                   |
| OLMo-2-1B                   | OLMo    | 1.5B   | 0.695        | [.679, .711] | 0.781                   |
| Qwen3-8B                    | Qwen    | 8.2B   | 0.692        | [.674, .708] | 0.714                   |
| Qwen3-0.6B                  | Qwen    | 0.6B   | 0.691        | [.673, .709] | 0.780                   |
| Mistral-7B                  | Mistral | 7.0B   | 0.633        | [.617, .648] | 0.682                   |
| <b>Average (8-gen LOGO)</b> |         |        | <b>0.694</b> | [.688, .700] | <b>0.762</b>            |
| <b>LOFO (family-level)</b>  |         |        | 0.686        | —            | —                       |

**Error analysis.** The gen-1 / gen-2+ boundary is the primary failure point: depth-2 samples (the shallowest within the 2+ class) exhibit the highest error rate (48.4%), decreasing monotonically to 20.3% at depth-4 (Appendix l) due to insufficient accumulated drift beyond the large gen-0→gen-1 distributional shift. Error rates decrease monotonically at deeper generations ( $\leq 29.2\%$  at gen-3+); DDR signal thus strengthens with chain length.

## 4.2 Differential Depth Response

**Formal definition.** For feature  $f$  measured across depths  $d \in \{0, 1, 2+\}$ , we define its *depth sensitivity* as  $\text{DDR}(f) = |\rho(\text{depth}, f)|$ , where  $\rho$  is Spearman correlation over all per-text measurements. A feature is classified as *drifting* if Cohen’s  $d > 0.50$  between adjacent depths, or *stationary* if  $d < 0.15$ ; the two-cluster structure is robust across  $\varepsilon \in [0.10, 0.75]$  ( $k$ -means silhouette = 0.660,  $p=0.030$ ; exhaustive split enumeration ranks this partition first; independent replication in all eight generators; Appendix v). Recursive inputs grow increasingly self-typical (Geng et al., 2026); this predicts location drift with shape invariance (confirmed across nucleus, greedy, and top- $k$  strategies; Appendix z.6).

Figure 1 reveals the dichotomy: central-tendency features (entropy mean, TTR, hapax ratio) decline monotonically with depth, while shape features (entropy max, tail decay, entropy min) remain flat (per-feature distributions: Appendix o; qualitative chain example: Appendix x). The partition is binary ( $k=2$ ; silhouette = 0.660), with entropy std as the sole *intermediate* member ( $d=0.46$ ,  $|\rho|=0.16$ ), distinct from three near-stationary features ( $d < 0.10$ ; Appendix v).

Pooled Spearman correlations corroborate the

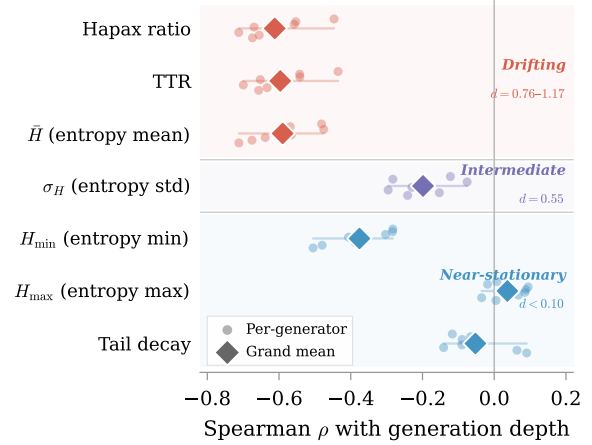


Figure 4: Differential depth response (DDR). Drifting features (red;  $|\rho| > 0.60$ , effect size  $d=0.74\text{--}1.05$ ) shift monotonically with generation depth; near-stationary features (blue;  $d \leq 0.10$ ) are approximately preserved; entropy std (purple;  $d=0.46$ ) is intermediate. Diamonds: cross-generator means; dots: per-generator values.

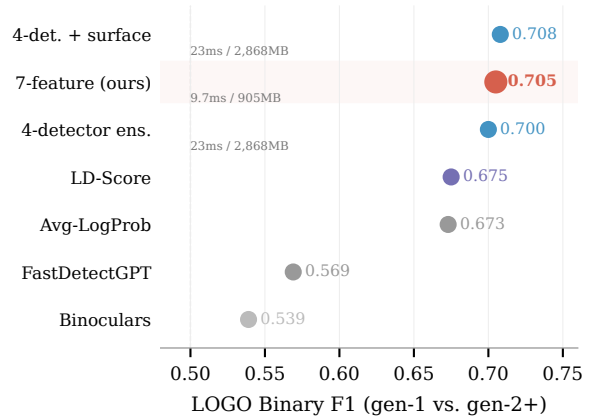


Figure 5: LOGO binary F1 vs. inference cost (8-gen pooled). The 7-feature ensemble (one 410M model) matches multi-detector ensembles (1.4B total) at  $\sim 2.4\times$  lower cost. RTX 5090 timings.

split: drifting features show  $|\rho| > 0.60$  (hapax ratio  $-0.64$ , TTR  $-0.62$ , entropy mean  $-0.60$ ) vs.  $|\rho| \leq 0.16$  for entropy max and tail decay; entropy min reaches  $|\rho|=0.40$  due to its gen-0/gen-1 boundary sensitivity (discussed below). This persistent magnitude gap defines the DDR partition (Figure 4).

**Quantification.** Cohen’s  $d$  shows a  $\geq 8\times$  gap between drifting ( $d=0.74\text{--}1.05$ ) and near-stationary ( $d < 0.10$ ); KS, Glass’s  $\Delta$ , Cliff’s  $\delta$  corroborate (Appendix v). Entropy min and tail decay are *boundary features*: entropy min is near-stationary for gen-1/gen-2+ ( $d=0.016$ ) but sensitive at gen-0/gen-1; tail decay passes two one-sided tests

(TOST) equivalence (Appendix u) for only 4/8 generators (per-generator  $|d|=0.07\text{--}0.22$ ).

**Validation.** Each hold-out generator independently recovers the identical partition (silhouette 0.505–0.760); TOST confirms near-stationarity for entropy min (8/8) and entropy max (6/8); permutation testing confirms the split as non-random ( $p=0.030$ ). Per-feature convergence dynamics (Figure 6), per-family DDR consistency (Appendix z.2), and causal feature analysis (Appendix z.4) appear in Appendices v–z.3.

**Feature complementarity.** The full 7-feature set ( $F1=0.771$ , 10-fold stratified cross-validation) outperforms surface-only (0.688) and reference-only (0.713; McNemar  $p \approx 0$ ; Appendix q). Group ablation (10-fold stratified CV) confirms complementarity: 3-class F1 gains +5.3 pp when non-drifting features are added (drifting-only 0.716, all-7 0.769), while binary detection is already saturated by location drift alone (drifting-only 0.697 vs. all-7 0.700); shape stability thus contributes to depth *grading*, not detection. TTR and hapax ratio ( $r = 0.97$ ) are retained for marginal LOFO gains (Appendix f).

**Cross-family reference models.** Alternative references (OLMo-2-1B, Qwen3-0.6B) yield F1 spread of only 0.013 (Appendix s); same-family proximity may inflate Pythia F1 by up to 1.2 pp, but the DDR partition is preserved.

**Temperature and prompt-type sensitivity.** All four tested generators degrade at  $t=0.9$  (−9.3 to −26.0 pp binary F1; Appendix h); continuation chains show the strongest drift ( $F1=0.781$ ; Appendix c). The DDR partition hierarchy is preserved across temperatures (Appendix z.1), including cross-temperature transfer (Appendix p); random word replacement at 10% degrades binary F1 by only 8 pp (Appendix i).

### 4.3 Comparison with Detection Baselines

We next ask whether existing detection methods, designed for binary real-vs-synthetic classification, incidentally capture depth structure, repurposing four detection scores as ordinal depth predictors under identical XGBoost and LOGO conditions (Table 3). The strongest baseline is a 4-detector + surface feature ensemble (binary  $F1=0.708$ ), which pools four detection scores with TTR and hapax ratio; among individual detectors, Avg-LogProb provides the most meaningful single-feature comparison.

Table 3: Depth classification (8-gen LOGO, pooled). Binary F1 excludes gen-0. 7-feature: one 410M model; 4-det ensembles: two models, 1.4B. QLoRA (§4.1) uses full token sequences (within-model 5-fold CV, not LOGO), serving as an information-asymmetric ceiling.

| Method           | Feat. | 3-class F1   | Bin F1       | ms/text    |
|------------------|-------|--------------|--------------|------------|
| 4-det. + surface | 6     | <b>0.795</b> | <b>0.708</b> | 22.8       |
| 4-detector ens.  | 4     | 0.785        | 0.700        | 22.8       |
| 7-feature (ours) | 7     | 0.774        | 0.705        | <b>9.7</b> |
| Avg-LogProb      | 1     | 0.707        | 0.673        | —          |
| FastDetectGPT    | 1     | 0.660        | 0.569        | —          |
| LD-Score         | 1     | 0.630        | 0.675        | —          |
| Binoculars       | 1     | 0.411        | 0.539        | —          |

The 7-feature ensemble matches the 4-detector+surface ensemble on binary F1 (0.705 vs. 0.708; Figure 5) at  $\sim 2.4\times$  lower cost, though the latter leads by 2.1 pp on 3-class F1 (0.795 vs. 0.774); both surpass individual detectors by 3.0–16.6 pp. The DDR decomposition explains why depth discrimination is achievable: location features carry the signal and shape features enable cross-generator transfer (§5), all with a single 410M reference model; ordinal-specific methods (CORAL, Frank-Hall) yield comparable F1 (Appendix h). Among individual detectors, LD-Score and Avg-LogProb best capture depth structure (binary  $F1=0.675$  and 0.673); FastDetectGPT captures 3-class ordering ( $F1=0.660$ ) but struggles with binary depth separation ( $F1=0.569$ ); Binoculars is least informative (binary  $F1=0.539$ ). Concurrent iterative-generation studies use different chain operators and cannot be directly included as depth baselines (Appendix z.8).

## 5 Discussion

DDR reflects monotonic NLL compression under recursive generation (Appendix z.9). Shape stability is not trivially predictable from architecture: entropy\_max is bounded by  $\log|V|\approx 10.8$  nats but observed values fall far below this ceiling; tail\_decay lacks an obvious bound yet remains near-stationary ( $|d|=0.10\text{--}0.22$ ); entropy\_std drifts ( $d=0.46$ ) despite shared NLL derivation. ANOVA confirms depth explains  $>48\times$  more variance in location than shape features (Appendix z.5).

**Relationship to model collapse.** Model collapse erodes tail coverage via weight updates, and shape features *do* change under retraining (Shumailov et al., 2024; Dohmatob et al., 2025). DDR, by contrast, characterizes fixed-parameter inference

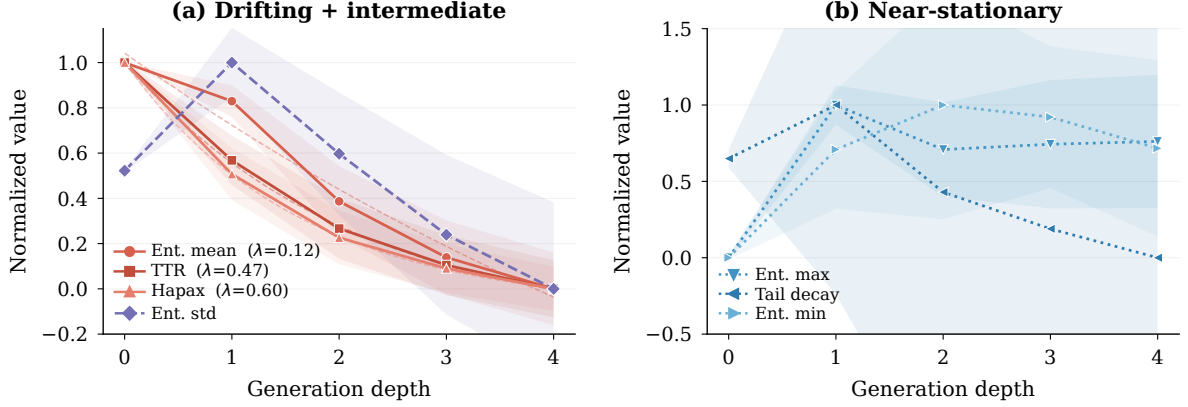


Figure 6: Feature convergence dynamics. (a) Drifting features (red) and intermediate entropy std (purple):  $|d| > 0.46$  by gen-2+. (b) Near-stationary features (blue):  $|d| < 0.25$  at all depths. Shape stability persists across chain lengths.

where shape remains stable. Shape stability is thus specific to the fixed-parameter regime, not an a priori consequence of architecture; DDR complements concurrent iterative-generation studies (Appendix z.8).

**Temperature sensitivity.** Higher sampling temperature ( $\tau > 0.7$ ) flattens token probability distributions, narrowing NLL variance and diluting the location shift that drifting features capture (§4.2). The DDR partition hierarchy is preserved across all tested temperatures (Appendix z.1), but signal strength is temperature-dependent, bounding current applicability to moderate sampling.

**Error structure and scope.** The gen-1→gen-2+ boundary is the primary failure point (error rate 48.4%; Appendix l): gen-1 texts retain sufficient diversity to overlap with gen-0; deeper generations show monotonically decreasing error ( $\leq 29.2\%$ ), so DDR-based classifiers are most reliable for heavily recycled content. The DDR partition is robust across decoding strategies, temperatures, and reference models (Appendices z.6–s). Instruction tuning compresses DDR (Appendix k), but QLoRA recovery to  $F1 = 0.89\text{--}0.91$  (within-model) shows the signal persists. Cross-domain ( $F1 = 0.74\text{--}0.75$ ; Appendix j) and 9-generator (Appendix z.7) evaluations confirm generalization. DDR enables corpus contamination detection and synthetic data pipeline auditing; depth-aware filtering does not improve perplexity (Appendix m), so DDR identifies *provenance* rather than *fluency degradation*. Finer-grained 5-class classification yields  $F1 \approx 0.50$  (Appendix d), a resolution limit of three depth classes.

## 6 Conclusion

Recursive generation produces a decomposable depth response: location features drift monotonically while shape features remain approximately preserved. This asymmetry supports cross-generator depth classification ( $F1 = 0.694\text{--}0.762$ ; 5-generator cross-model  $F1 = 0.830$ ) under moderate sampling temperatures ( $\tau \leq 0.7$ ). Instruction tuning compresses DDR, bounding current applicability to base models; QLoRA recovery ( $F1 = 0.89\text{--}0.91$ , within-model) points to tractable extension to instruct settings.

## Limitations

All eight generators are base (pretrained) models with  $\leq 8.2B$  parameters; instruction-tuned variants show substantially degraded depth signal (within-model 3-class  $F1$  drops 3.7–19.9 pp; Appendix k), though QLoRA partially recovers ( $F1 = 0.89\text{--}0.91$ , within-model); this pattern is consistent with signal compression rather than elimination. The evaluation assumes English Wikipedia seeds, a fixed decoding regime ( $t=0.7$ , nucleus  $p=0.9$ ), and an English-centric reference model (Pythia-410M); cross-temperature transfer is tested (Appendix p) but cross-lingual generalization remains untested. Depth resolution is bounded at three classes (gen-0 / gen-1 / gen-2+); finer-grained 5-class classification yields  $F1 \approx 0.50$  with deeper depths mutually indistinguishable (Appendix d). DDR signal strength is temperature-dependent; at  $\tau=0.9$ , classification performance degrades substantially (−9.3 to −26.0 pp binary  $F1$ ), and cross-temperature transfer at production scales remains untested.

## Ethics Statement

All texts analyzed in this work are LLM-generated from publicly available Wikipedia seeds; no personal or sensitive data is involved. Our method estimates the generation depth of a text passage (how many recursive generation steps it has undergone), not the identity of its author. Individual authorship attribution is outside both the scope and the capability of this approach; privacy implications are minimal. We acknowledge that depth estimation tools could in principle be used to selectively suppress machine-generated text in contexts where such content is legitimate (e.g., accessibility tools, assistive writing). We encourage deployment in data governance pipelines where provenance transparency, rather than content removal, is the objective. All code, data, and trained models will be released upon publication.

**Licenses.** All generator models are publicly available under permissive licenses (Apache 2.0 for Pythia, OLMo, and Mistral; Qwen License for Qwen3; Llama Community License for Llama 3.2). Wikipedia seed passages are distributed under CC BY-SA 3.0.

**Compute budget.** All experiments were conducted on 4× NVIDIA RTX 5090 GPUs (32 GB each), totaling approximately 62 GPU-hours.

**Software.** Experiments used Python 3.12.3, PyTorch 2.10.0 (CUDA 12.8), Transformers 4.57.6, XGBoost 3.2.0, scikit-learn 1.8.0, NumPy 2.2.6, pandas 3.0.2, and SciPy 1.17.1.

## References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoobi, and Richard G. Baraniuk. 2024. Self-Consuming Generative Models Go MAD. In *Proceedings of the International Conference on Learning Representations*.

Mohd Ruhul Ameen, Akif Islam, Nadim Mahmud, and Md. Ekramul Hamid. 2026. Chainwash: Multi-Step Rewriting Attacks on Diffusion Language Model Watermarks. *arXiv preprint arXiv:2605.05503*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *Proceedings of the International Conference on Learning Representations*.

Advik Raj Basani and Pin-Yu Chen. 2026. *Diversity Boosts AI-Generated Text Detection*. *Transactions on Machine Learning Research*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the International Conference on Machine Learning*.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation. *Pattern Recognition Letters*, 140:325–331.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zihao Cheng, Li Zhou, Feng Jiang, Benyou Wang, and Haizhou Li. 2025. Beyond Binary: Towards Fine-Grained LLM-Generated Text Detection via Role Recognition and Involvement Measurement. In *Proceedings of the ACM Web Conference*.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Posen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Mery, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, and 5 others. 2024. Scalable Watermarking for Identifying Large Language Model Outputs. *Nature*, 634:818–823.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Fine-tuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. 2025. Strong Model Collapse. In *Proceedings of the International Conference on Learning Representations*.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. 2024. A Tale of Tails: Model Collapse as a Change of Scaling Laws. In *Proceedings of the International Conference on Machine Learning*.

Eibe Frank and Mark Hall. 2001. A Simple Approach to Ordinal Classification. In *Proceedings of the European Conference on Machine Learning*.

Mingmeng Geng, Amr Mohamed, Guokan Shang, Michalis Vazirgiannis, and Thierry Poibeau. 2026. Markovian Generation Chains in Large Language Models. *arXiv preprint arXiv:2603.11228*.

|     |                                                            |                                                             |     |
|-----|------------------------------------------------------------|-------------------------------------------------------------|-----|
| 656 | Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,        | Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley             | 714 |
| 657 | Abhinav Pandey, Abhishek Kadian, Ahmad Al-                 | Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao,              | 715 |
| 658 | Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,         | Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christo-          | 716 |
| 659 | Amy Yang, and 1 others. 2024. The Llama 3 Herd of          | pher Potts, Christopher D. Manning, and James Y.            | 717 |
| 660 | Models. <i>arXiv preprint arXiv:2407.21783</i> .           | Zou. 2024b. Mapping the Increasing Use of LLMs in           | 718 |
| 661 | Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bha-      | Scientific Papers. In <i>Proceedings of the Conference</i>  | 719 |
| 662 | gia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh           | on Language Modeling.                                       | 720 |
| 663 | Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang,           | Eric Mitchell, Yoonho Lee, Alexander Khazatsky,             | 721 |
| 664 | Shane Arora, David Atkinson, Russell Authur, Khy-          | Christopher D. Manning, and Chelsea Finn. 2023.             | 722 |
| 665 | athi Raghavi Chandu, Arman Cohan, Jennifer Dumas,          | DetectGPT: Zero-Shot Machine-Generated Text De-             | 723 |
| 666 | Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others.       | tection using Probability Curvature. In <i>Proceedings</i>  | 724 |
| 667 | 2024. OLMo: Accelerating the Science of Language           | of the International Conference on Machine Learn-           | 725 |
| 668 | Models. In <i>Proceedings of the Annual Meeting of the</i> | ing.                                                        | 726 |
| 669 | <i>Association for Computational Linguistics</i> .         |                                                             |     |
| 670 | Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis,           | Amr Mohamed, Mingmeng Geng, Michalis Vazirgian-             | 727 |
| 671 | and Chloé Clavel. 2024. The Curious Decline of             | nis, and Guokan Shang. 2025. LLM as a Broken                | 728 |
| 672 | Linguistic Diversity: Training Language Models on          | Telephone: Iterative Generation Distorts Information.       | 729 |
| 673 | Synthetic Text. In <i>Findings of the Association for</i>  | In <i>Proceedings of the Annual Meeting of the Associa-</i> | 730 |
| 674 | <i>Computational Linguistics: NAACL</i> .                  | <i>tion for Computational Linguistics</i> .                 | 731 |
| 675 | Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier         | Priyadarshan Narayanasamy, Swastik Agrawal, Klint           | 732 |
| 676 | Sánchez-Monedero, Francisco Fernández-Navarro,             | Faber, and Fardina Fathmiul Alam. 2026. Beyond              | 733 |
| 677 | and César Hervás-Martínez. 2016. Ordinal Re-               | Perplexity: Character Distribution Signatures and           | 734 |
| 678 | gression Methods: Survey and Experimental Study.           | the MDTA Benchmark for AI Text Detection. <i>arXiv</i>      | 735 |
| 679 | <i>IEEE Transactions on Knowledge and Data Engi-</i>       | <i>preprint arXiv:2605.01647</i> .                          | 736 |
| 680 | <i>neering</i> , 28(1):127–146.                            |                                                             |     |
| 681 | Abhimanyu Hans, Avi Schwarzschild, Valeriia                | Vishakh Padmakumar and He He. 2024. Does Writing            | 737 |
| 682 | Cherepanova, Hamid Kazemi, Aniruddha Saha,                 | with Language Models Reduce Content Diversity?              | 738 |
| 683 | Micah Goldblum, Jonas Geiping, and Tom Goldstein.          | In <i>Proceedings of the International Conference on</i>    | 739 |
| 684 | 2024. Spotting LLMs With Binoculars: Zero-Shot             | <i>Learning Representations</i> .                           | 740 |
| 685 | Detection of Machine-Generated Text. In <i>Proceed-</i>    |                                                             |     |
| 686 | <i>ings of the International Conference on Machine</i>     | Guilherme Penedo, Hynek Kydlíček, Loubna Ben Al-            | 741 |
| 687 | <i>Learning</i> .                                          | lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel,        | 742 |
| 688 | Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-       | Leandro Von Werra, and Thomas Wolf. 2024. The               | 743 |
| 689 | sch, Chris Bamford, Devendra Singh Chaplot, Diego          | FineWeb Datasets: Decanting the Web for the Finest          | 744 |
| 690 | de las Casas, Florian Bressand, Gianna Lengyel, Guil-      | Text Data at Scale. In <i>Advances in Neural Informa-</i>   | 745 |
| 691 | laume Lample, Lucile Saulnier, and 1 others. 2023.         | <i>tion Processing Systems: Datasets and Benchmarks</i>     | 746 |
| 692 | Mistral 7B. <i>arXiv preprint arXiv:2310.06825</i> .       | <i>Track</i> .                                              | 747 |
| 693 | John Kirchenbauer, Jonas Geiping, Yuxin Wen,               | Qwen Team. 2025. Qwen3 Technical Report. <i>arXiv</i>       | 748 |
| 694 | Jonathan Katz, Ian Miers, and Tom Goldstein. 2023.         | <i>preprint arXiv:2505.09388</i> .                          | 749 |
| 695 | A Watermark for Large Language Models. In <i>Pro-</i>      | Zixin Rao, Youssef Mohamed, Shang Liu, and Zeyan            | 750 |
| 696 | <i>ceedings of the International Conference on Machine</i> | Liu. 2025. Two Birds with One Stone: Multi-Task             | 751 |
| 697 | <i>Learning</i> .                                          | Detection and Attribution of LLM-Generated Text.            | 752 |
| 698 | Ryuto Koike, Liam Dugan, Masahiro Kaneko, Chris            | <i>arXiv preprint arXiv:2508.14190</i> .                    | 753 |
| 699 | Callison-Burch, and Naoaki Okazaki. 2025. Machine          | Vinu Sankar Sadasivan, Aounon Kumar, Sriram Bala-           | 754 |
| 700 | Text Detectors are Membership Inference Attacks.           | subramanian, Wenxiao Wang, and Soheil Feizi. 2025.          | 755 |
| 701 | <i>arXiv preprint arXiv:2510.19492</i> .                   | Can AI-Generated Text be Reliably Detected? Stress          | 756 |
| 702 | Daniël Lakens. 2017. Equivalence Tests: A Prac-            | Testing AI Text Detectors Under Various Attacks.            | 757 |
| 703 | tical Primer for $t$ Tests, Correlations, and Meta-        | <i>Transactions on Machine Learning Research</i> .          | 758 |
| 704 | Analyses. <i>Social Psychological and Personality Sci-</i> | Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu,       | 759 |
| 705 | <i>ence</i> , 8(4):355–362.                                | and Sanmi Koyejo. 2025. Position: Model Collapse            | 760 |
| 706 | Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp,      | Does Not Mean What You Think. <i>arXiv preprint</i>         | 761 |
| 707 | Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Hao-           | <i>arXiv:2503.03150</i> .                                   | 762 |
| 708 | tian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland,        | Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo           | 763 |
| 709 | and James Y. Zou. 2024a. Monitoring AI-Modified            | Huang, Daogao Liu, Terra Blevins, Danqi Chen, and           | 764 |
| 710 | Content at Scale: A Case Study on the Impact of            | Luke Zettlemoyer. 2024. Detecting Pretraining Data          | 765 |
| 711 | ChatGPT on AI Conference Peer Reviews. In <i>Pro-</i>      | from Large Language Models. In <i>Proceedings of the</i>    | 766 |
| 712 | <i>ceedings of the International Conference on Machine</i> | <i>International Conference on Learning Representa-</i>     | 767 |
| 713 | <i>Learning</i> .                                          | <i>tions</i> .                                              | 768 |

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2023. Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities. *Pattern Analysis and Applications*, 26:941–955.

Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. SemEval-2024 Task 8: Multidomain, Multimodal and Multilingual Machine-Generated Text Detection. In *Proceedings of the International Workshop on Semantic Evaluation*.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. 2025. [Unveiling Attractor Cycles in Large Language Models: A Dynamical Systems View of Successive Paraphrasing](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Language Resources and Evaluation Conference*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Computational Linguistics*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data Selection for Language Models via Importance Resampling. In *Advances in Neural Information Processing Systems*.

Table 4: Per-generator sample counts after filtering (8 generators, 5 families). Each generator starts from ~1,000 shared Wikipedia seeds and produces chains of depth 4 (5 texts per chain). Attrition varies by generator due to differing generation failure rates.<sup>3</sup>

| Generator       | Family  | Params | Samples       |
|-----------------|---------|--------|---------------|
| Pythia-1B       | Pythia  | 1.0B   | 4,555         |
| Pythia-2.8B     | Pythia  | 2.8B   | 4,609         |
| OLMo-2-1B       | OLMo    | 1.5B   | 4,713         |
| Qwen3-0.6B      | Qwen    | 0.6B   | 4,538         |
| Qwen3-1.7B      | Qwen    | 1.7B   | 4,617         |
| Qwen3-8B        | Qwen    | 8.2B   | 4,497         |
| Llama-3.2-3B    | Llama   | 3.2B   | 4,680         |
| Mistral-7B-v0.3 | Mistral | 7.0B   | 4,849         |
| <b>Total</b>    |         |        | <b>37,058</b> |

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. [Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2024. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. In *Proceedings of the International Conference on Learning Representations*.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. 2025. [Human Texts Are Outliers: Detecting LLM-generated Texts via Out-of-distribution Detection](#). In *Advances in Neural Information Processing Systems*.

**Note:** Unless otherwise stated, appendix experiments use the original six-generator subset (Pythia-2.8B/1B, OLMo-2-1B, Qwen3-1.7B/0.6B/8B). Main paper results (§4) use the full eight-generator set (adding Llama-3.2-3B and Mistral-7B-v0.3).

## a Dataset Statistics

Table 4 reports per-generator sample counts after filtering (removing generation failures, exact duplicates, and passages shorter than 50 whitespace tokens). Under the 3-class evaluation scheme, the dataset contains 7,963 gen-0 samples (21.5%), 7,896 gen-1 (21.3%), and 21,199 gen-2+ (57.2%, combining depths 2–4). The gen-2+ majority reflects the chain structure (three deeper depths per chain vs. one each for gen-0 and gen-1); class-balanced sampling is applied during evaluation.

<sup>3</sup>Table 3 evaluates on a subset of 34,131 samples for which all detector scores (FastDetectGPT, Binoculars, LD-Score, Avg-LogProb) are available.

Figure 7 shows the full distribution of each feature across five generation depths (0–4), pooled across all eight generators ( $N = 37,058$ ). The violin plots confirm the DDR pattern: drifting features (entropy mean, TTR, hapax ratio) show progressive narrowing and location shift, while shape features (entropy max, tail decay) maintain approximately constant distributions across depths.

## b Length Confound Analysis

Surface diversity features (TTR, hapax ratio) mechanically depend on text length: shorter texts yield higher type-token ratios by construction. If generation depth systematically shortens text, observed depth trends could be length artifacts rather than genuine distributional shifts. We investigate with three complementary analyses.

**(a) Length–depth correlation.** Spearman rank correlation between word count and generation depth is weak across all eight generators ( $|\rho| < 0.14$ ); depth does not systematically shorten or lengthen text in our corpus.

**(b) Length–feature correlations.** While length–depth correlation is weak, individual features show moderate length dependence: tail\_decay ( $|\rho| = 0.38$ ), hapax\_ratio ( $|\rho| = 0.30$ ), and TTR ( $|\rho| = 0.29$ ). Reference-model features show weaker length correlations (entropy\_mean  $|\rho| = 0.09$ , entropy\_std  $|\rho| = 0.18$ ). This motivates the length-matched evaluation below: if length mediates some depth–feature association, controlling for length will reduce classification performance.

**(c) Length-matched classification.** We subsample the dataset so that word-count distributions are equalized across depth classes (gen-0/gen-1/gen-2+), then retrain the binary XGBoost classifier under pooled LOGO (8-fold). Length-matched binary F1 = 0.724, down from 0.770 on the unmatched set ( $\Delta = -0.046$ ). The matched subset is smaller than the full corpus; a random-subsample control at equal  $N$  (below) isolates the deconfounding effect from the sample-size reduction. To distinguish this from a sample-size effect, we run a random-subsample control at equal  $N$  without length matching: F1 = 0.767, which confirms that the 0.046 drop reflects genuine length deconfounding rather than reduced data.

**Conclusion.** Length accounts for approximately 6% of discriminative performance ( $\Delta = 0.046$  out

of 0.770). The majority of the depth signal is length-independent: length-matched F1 (0.724) remains far above chance and substantially above all individual detection baselines (§4.3). Surface features (TTR, hapax\_ratio) partially co-vary with length, but their depth-discriminative contribution persists after length control, consistent with generation-induced vocabulary compression operating independently of text truncation.

## c Prompt Type Generalization

We evaluate depth profiling under three prompt types using Pythia-2.8B. Each uses 995 seeds  $\times$  5 depths. The exact templates are:

### Prompt Templates

**Continuation:** Continue the following text: {text}  
**Elaboration:** Elaborate on the following text: {text}  
**Inspired rewriting:** Write a new passage inspired by the following text: {text}

**Within-prompt-type classification.** 5-fold CV macro-F1: continuation 0.78, rewriting 0.74, elaboration 0.73. All three prompt types exhibit the same differential depth response pattern: drifting features (hapax ratio, TTR, entropy mean) show consistent monotonic trends, while near-stationary features (entropy max, tail decay) change by less than 5%.

**Cross-prompt-type transfer (LOPTO).** Leave-one-prompt-type-out evaluation yields average F1 = 0.753, which confirms that depth signal transfers across prompt types. Feature decay rates differ: continuation shows the strongest drift (16.9% per depth step), followed by rewriting (6.1%) and elaboration (1.4%), consistent with continuation chains applying less semantic constraint on the output.

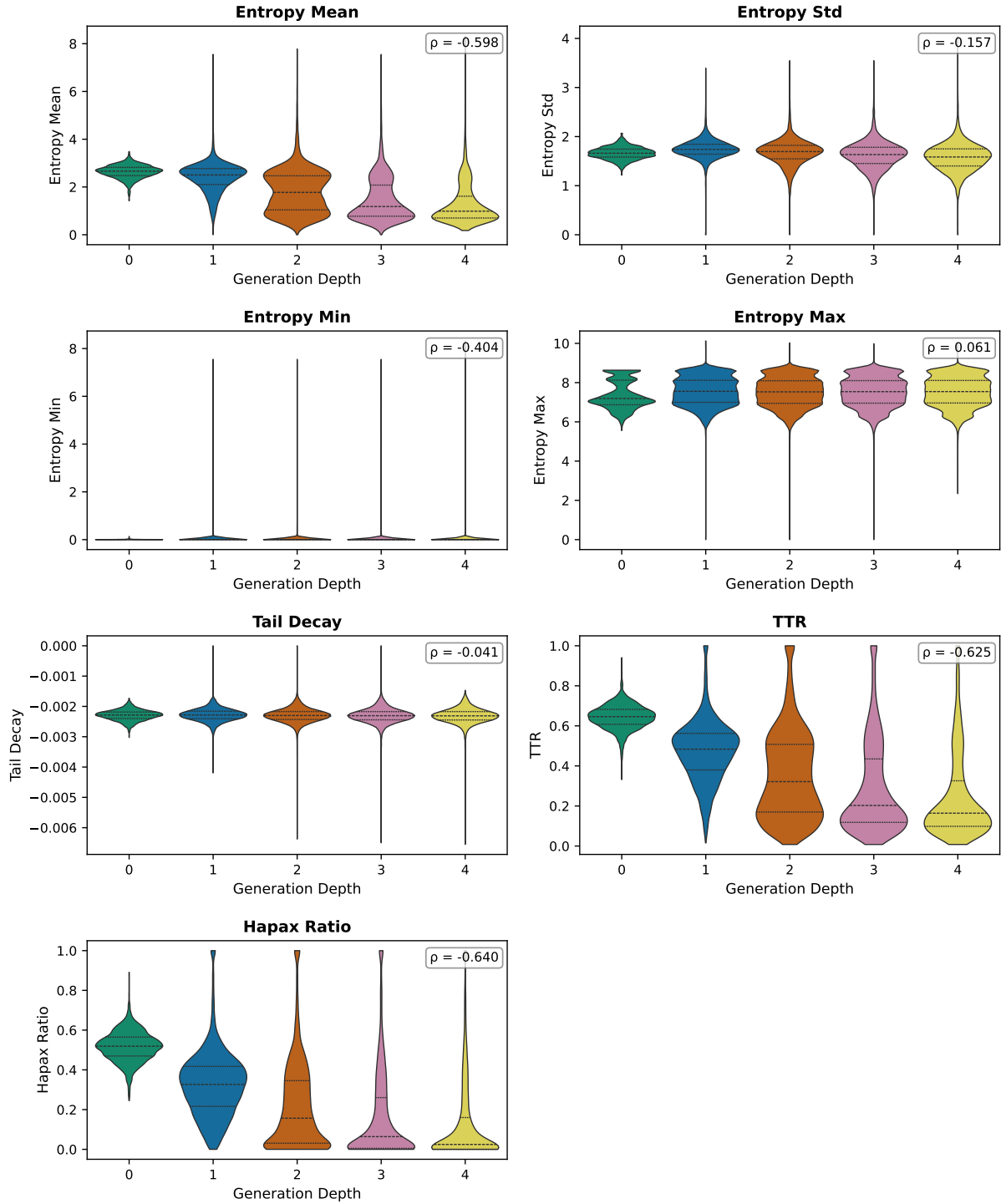


Figure 7: Feature distributions by generation depth (violin plots,  $N = 37,058$ ). Drifting features (entropy mean, TTR, hapax ratio:  $|\rho| > 0.60$ ) show clear location shifts and narrowing with increasing depth. Shape features (entropy max, tail decay:  $|\rho| < 0.10$ ) remain approximately stationary. Quartile lines indicate median and interquartile range.

## d Finer-Grained Depth Classification

We evaluate finer-grained 5-class depth classification (gen-0 through gen-4) using the same 7-feature ensemble and LOGO protocol described in §3.

Table 5: 5-class LOGO evaluation. Each row holds out one generator for testing, training on the remaining five.

| Held-out Generator | Macro F1             | Weighted F1  | Accuracy     |
|--------------------|----------------------|--------------|--------------|
| OLMo-2-1B          | 0.484                | 0.506        | 0.524        |
| Pythia-1B          | 0.543                | 0.568        | 0.575        |
| Pythia-2.8B        | 0.523                | 0.540        | 0.560        |
| Qwen3-0.6B         | 0.513                | 0.521        | 0.546        |
| Qwen3-1.7B         | 0.510                | 0.536        | 0.548        |
| Qwen3-8B           | 0.434                | 0.455        | 0.493        |
| <b>Average</b>     | <b>0.501 ± 0.038</b> | <b>0.521</b> | <b>0.541</b> |

Table 6: Per-class F1 averaged across all six LOGO folds.

| Metric | Gen-0 | Gen-1 | Gen-2 | Gen-3 | Gen-4 |
|--------|-------|-------|-------|-------|-------|
| Avg F1 | 0.921 | 0.634 | 0.268 | 0.259 | 0.423 |

**Analysis.** Moving from 3-class (gen-0 / gen-1 / gen-2+) to 5-class classification reduces LOGO macro F1 to approximately 0.50 (six-generator subset; the 3-class baseline for this subset is comparable to the 8-generator result of 0.762 in Table 2). The degradation is concentrated in the middle depths: gen-2 and gen-3 achieve only 0.268 and 0.259 F1, respectively. The aggregated confusion matrix reveals that 36.5% of true gen-2 samples are misclassified as gen-1, while 37.8% of true gen-3 samples are predicted as gen-4. Adjacent depths (gen-2 ↔ gen-3, gen-3 ↔ gen-4) account for the majority of errors, consistent with the intuition that deeper recursive generations converge in surface-level statistical signatures. Gen-0 (human text) remains highly distinguishable (F1 = 0.921), and gen-1 retains moderate separability (F1 = 0.634). Shape features that are near-stationary in the 3-class setting become substantially more important under 5-class evaluation: permutation importance of entropy\_std increases by 93% and tail\_decay by 400% relative to 3-class, yet these gains are insufficient to resolve adjacent depths. These results support the 3-class grouping adopted in the main paper: collapsing gen-2+ preserves the most informative boundary (human vs. first-generation vs. deeper recursion) while avoiding the noise introduced by near-identical adjacent depths.

**Classifier comparison.** To verify that 5-class degradation reflects feature-level limitations rather

than classifier choice, we evaluate four alternatives under the same protocol: HistGradientBoosting (macro F1 = 0.501, 95% bootstrap CI [0.496, 0.506]), random forest (0.476), logistic regression (0.447), and Frank-Hall ordinal reduction (0.451). All classifiers perform comparably to the default XGBoost ensemble (0.501); classifier capacity is not the bottleneck. The gen-2/gen-3 boundary is the primary failure mode across all methods: these two classes are effectively inseparable with the current feature set. This establishes 3-class granularity as the practical ceiling for surface and entropy features. At the 3-class level, ordinal-specific methods (CORAL ordinal regression, Frank-Hall reduction) similarly fail to improve over multiclass XGBoost (§4), which confirms that the bottleneck is feature separability across all granularities, not the modeling of label ordinality.

## e End-to-End Depth Classification (QLoRA)

To assess whether depth signal extends beyond handcrafted features, we fine-tune a pretrained language model end-to-end as a depth classifier using QLoRA (Detrmers et al., 2023).

**Setup.** We fine-tune Qwen3-0.6B-Base with LoRA rank 16, learning rate  $2 \times 10^{-4}$ , and 3 epochs, framing depth estimation as 5-class sequence classification (gen-0 through gen-4).<sup>4</sup> For 3-class evaluation, we collapse predictions into gen-0 / gen-1 / gen-2+ post hoc. All results use 5-fold stratified cross-validation on the 4-generator subset.

Table 7: QLoRA end-to-end depth classification vs. XGBoost on handcrafted features, under 5-fold stratified CV (4-generator subset).

| Method               | 5-class F1    | 3-class F1    | Per-class F1 (5-class)                |
|----------------------|---------------|---------------|---------------------------------------|
| XGBoost (7 features) | 0.497 ± 0.008 | 0.805 ± 0.008 | —                                     |
| QLoRA (Qwen3-0.6B)   | 0.583 ± 0.010 | 0.905 ± 0.003 | 0.999 / 0.796 / 0.413 / 0.259 / 0.447 |

**Results.** QLoRA achieves 3-class F1 = 0.905 (±0.003) under 5-fold stratified CV on a 4-generator subset, a +10.0 pp improvement over the same-subset XGBoost baseline (0.805). On 5-class evaluation, QLoRA reaches F1 = 0.583 (±0.010) vs. 0.497 for XGBoost (+8.6 pp), with the largest gains at gen-1 (0.796 vs. 0.650) and gen-2 (0.413

<sup>4</sup>Evaluated on a 4-generator subset (Pythia-1B, OLMo-2-1B, Qwen3-0.6B, Qwen3-1.7B) using 5-fold stratified CV; XGBoost 3-class F1 = 0.805 under the same protocol. We expect 8-generator training to further improve given additional diversity; the comparison direction is preserved.

vs. 0.267). The gen-2/gen-3 boundary remains the primary failure mode (gen-3 F1 = 0.259), consistent with the feature-level analysis in §5.

**Hyperparameter robustness.** A second run with LoRA rank 32 and seed 123 yields 5-class F1 = 0.587 ( $\pm 0.005$ ) and 3-class F1 = 0.905 ( $\pm 0.004$ ), which confirms stability across hyperparameter and seed variation ( $\Delta < 1$  pp on both metrics).

**Implications.** The substantial 3-class improvement demonstrates that raw token embeddings contain depth signal beyond what summary statistics capture; the seven handcrafted features are a strong but lossy compression of the full distributional information. End-to-end fine-tuning exploits token-level patterns (e.g., positional regularities, n-gram repetitions) that aggregate features cannot represent. This confirms that the depth signal documented in our main analysis is real and exploitable beyond engineered features, while the handcrafted approach retains practical advantages: interpretability, computational efficiency (no GPU inference), and the ability to reveal the differential depth response pattern that motivates the task formulation. The QLoRA ceiling thus represents an information-asymmetric upper bound rather than a competing method: it consumes the full token sequence (0.6B parameters, GPU inference) whereas the 7-feature ensemble distills the signal into interpretable scalars that expose *why* depth is detectable, a decomposition that end-to-end models cannot provide.

## f Leave-One-Feature-Out Analysis

We complement the permutation importance analysis (§4.2) with a leave-one-feature-out ablation. Each row drops one of the seven features and re-evaluates the full 6-gen LOGO protocol;  $\Delta$  is the change relative to the full 7-feature ensemble (F1 = 0.789).

**Analysis.** Entropy\_max causes the largest degradation when removed ( $-1.5$  pp), despite ranking only fourth in permutation importance (split frequency 0.063). This indicates that entropy\_max provides unique signal (capturing the most surprising token under the reference model) that is not redundant with other features. Conversely, entropy\_min is fully redundant ( $\Delta = +0.01$  pp), consistent with its near-zero permutation importance (0.015). Six of seven features contribute

Table 8: Leave-one-feature-out ablation (6-gen LOGO). Each row drops one feature and re-evaluates with the LOGO protocol. Sorted by impact magnitude.

| Dropped feature   | LOGO F1 | $\Delta$ vs Full |
|-------------------|---------|------------------|
| entropy_max       | 0.774   | $-1.50$ pp       |
| entropy_mean      | 0.783   | $-0.63$ pp       |
| ttr               | 0.784   | $-0.55$ pp       |
| hapax_ratio       | 0.785   | $-0.42$ pp       |
| tail_decay        | 0.786   | $-0.33$ pp       |
| entropy_std       | 0.786   | $-0.29$ pp       |
| entropy_min       | 0.789   | $+0.01$ pp       |
| Full (7 features) | 0.789   | —                |

positively to the ensemble, which confirms that the 7-feature design is not over-specified.

To verify that feature importance generalizes beyond the original 6-generator subset, we repeat the LOFO analysis on the full 8-generator set (37,058 samples, 10-fold stratified CV). Table 9 confirms consistent rankings: entropy\_max remains the most impactful feature ( $\Delta = -0.019$ ), followed by entropy\_mean ( $\Delta = -0.012$ ). All seven features contribute positively, with no feature becoming redundant upon dataset expansion.

**Minimal sufficient subset.** To assess feature redundancy, we evaluate all  $\binom{7}{2}=21$  two-feature subsets under 10-fold stratified CV (8-gen, 3-class). The best pair is {entropy\_mean, entropy\_max}, achieving F1 = 0.68 (one drifting and one stationary feature), which captures 88% of the full-ensemble performance (0.771) and confirms that the DDR partition’s two components are individually informative. Adding the remaining five features yields +9.1 pp; the full set provides meaningful complementarity beyond the minimal core. Nevertheless, the two-feature result demonstrates that the depth signal is not an artifact of feature-set size.

Table 9: Leave-one-feature-out importance on the full 8-generator set (37,058 samples). Ranking is consistent with the 6-generator subset (Table 8).

| Configuration     | 3-class F1 | $\Delta$ |
|-------------------|------------|----------|
| Full (7 features) | 0.771      | —        |
| – entropy_max     | 0.752      | $-0.019$ |
| – entropy_mean    | 0.759      | $-0.012$ |
| – entropy_std     | 0.763      | $-0.007$ |
| – ttr             | 0.764      | $-0.007$ |
| – hapax_ratio     | 0.765      | $-0.005$ |
| – entropy_min     | 0.766      | $-0.005$ |
| – tail_decay      | 0.767      | $-0.004$ |

## f.1 Candidate Feature Screening

To verify that the seven retained features are not an arbitrary subset, we screened six additional NLL-distributional statistics as candidate features: median, 90th percentile (p90), interquartile range (IQR), coefficient of variation (CV), skewness, and kurtosis. Table 10 reports their DDR coordinates, classification impact, and correlation with retained features.

Table 10: Candidate feature screening. All six candidates fall in the drifting DDR zone and correlate strongly ( $|r| > 0.65$ ) with existing features. Adding any combination to the baseline degrades or leaves unchanged LOGO 3-class F1; LOFO  $\Delta > 0$  indicates that removing the candidate *improves* classification.

| Candidate | $ \rho $ | $d$   | LOFO $\Delta F1$ | $ r _{\max}$ | Most correlated |
|-----------|----------|-------|------------------|--------------|-----------------|
| median    | 0.685    | 0.794 | +0.008           | 0.960        | entropy_mean    |
| p90       | 0.693    | 0.960 | +0.000           | 0.937        | entropy_mean    |
| IQR       | 0.708    | 1.079 | +0.007           | 0.908        | entropy_std     |
| CV        | 0.633    | 0.980 | -0.000           | 0.761        | entropy_mean    |
| skewness  | 0.651    | 0.940 | +0.010           | 0.780        | entropy_mean    |
| kurtosis  | 0.635    | 0.769 | -0.003           | 0.650        | entropy_mean    |

All six candidates occupy the drifting zone ( $|\rho| > 0.63$ ,  $d > 0.77$ ), overlapping with entropy mean, TTR, and hapax ratio; none introduces a new axis of depth-dependent variation. Three candidates (median,  $r=0.96$ ; p90,  $r=0.94$ ; IQR,  $r=0.91$ ) are near-perfect linear proxies of existing features. Adding all six to the 6-feature baseline<sup>5</sup> reduces 3-class LOGO F1 from 0.733 to 0.722, consistent with overfitting from redundant inputs. LOFO analysis on the extended 12-feature set confirms that removing skewness (+0.010), median (+0.008), or IQR (+0.007) *improves* classification; these features inject noise rather than signal. The retained seven features span three complementary DDR zones (drifting, intermediate, near-stationary) and two modalities (reference-model NLL, surface lexical diversity); additional distributional statistics provide no incremental information.

## g Confusion Matrix

Table 11 reports the aggregated confusion matrix from the original six-generator subset (Pythia-2.8B, Pythia-1B, OLMo-2-1B, Qwen3-1.7B, Qwen3-0.6B, Qwen3-8B) under leave-one-generator-out (LOGO) evaluation, collapsed to three ordinal classes: gen-0 (human), gen-1 (first-generation),

and gen-2+ (depth  $\geq 2$ ). The later-added Llama-3.2-3B and Mistral-7B-v0.3 are evaluated in the full 8-generator LOGO results (Table 2).

Table 11: Aggregated 3-class confusion matrix from 6-generator LOGO evaluation ( $N = 27,529$ ). Each cell shows count and row-normalized percentage. Gen-0 is near-perfectly identified. Gen-1 errors are asymmetric: misclassification as gen-2+ (18.8%) is  $3\times$  more frequent than as gen-0 (6.2%). Gen-1 text already carries detectable recursive distortion.

|             | Pred gen-0    | Pred gen-1    | Pred gen-2+    |
|-------------|---------------|---------------|----------------|
| True gen-0  | 5,937 (99.4%) | 35 (0.6%)     | 0 (0.0%)       |
| True gen-1  | 368 (6.2%)    | 4,437 (75.0%) | 1,113 (18.8%)  |
| True gen-2+ | 464 (3.0%)    | 3,646 (23.3%) | 11,529 (73.7%) |

Figure 8 provides a heatmap visualization of the confusion matrix. The row-normalized view highlights near-perfect gen-0 classification (99.4% recall), gen-1 confusion primarily with gen-2+ (18.8%) rather than gen-0 (6.2%), and gen-2+ misclassification predominantly into gen-1 (23.3%), consistent with DDR predictions.

## h Classifier Sensitivity

To verify that the depth signal resides in the features rather than the classifier, we repeat the 6-generator LOGO evaluation using Logistic Regression and Random Forest in place of XGBoost (Table 12). All other settings (the same 7 features, 3-class labels, and class-balanced downsampling) are kept identical. For Logistic Regression we apply standard scaling to the features.

Table 12: LOGO macro-F1 across three classifiers. Each column shows the held-out generator F1. (6-generator LOGO; 3-class.)

| Classifier     | Avg F1 | OLMo-2 | Py-1B | Py-2.8B | Qw-0.6B | Qw-1.7B | Qw-8B |
|----------------|--------|--------|-------|---------|---------|---------|-------|
| XGBoost        | 0.789  | 0.764  | 0.856 | 0.804   | 0.824   | 0.806   | 0.683 |
| Random Forest  | 0.790  | 0.764  | 0.856 | 0.804   | 0.825   | 0.807   | 0.684 |
| Logistic Regr. | 0.725  | 0.719  | 0.781 | 0.740   | 0.745   | 0.743   | 0.624 |

Both tree-based classifiers achieve LOGO F1  $\approx 0.79$ , and even a linear model reaches 0.725, well above the 0.33 chance baseline for three classes. The per-generator ranking is also consistent across classifiers (Pythia-1B easiest, Qwen3-8B hardest), confirming that discriminative power originates in the entropy and lexical features, not in classifier capacity. Ordinal-specific methods (CORAL (Cao et al., 2020), Frank-Hall (Frank and Hall, 2001)) achieve comparable F1 with overlapping CIs, which confirms that 3-class granularity provides

<sup>5</sup>Tail decay requires full-vocabulary logits unavailable in the token-NLL subset used for this analysis; we report 6-feature baselines for fair comparison.

insufficient resolution for ranking-aware losses to improve over standard multiclass classification.

**XGBoost hyperparameter sensitivity.** To verify that performance is not an artifact of hyperparameter selection, we vary `max_depth`  $\in \{3, 4, 6, 8\}$  while keeping all other settings fixed (Table 13).

Table 13: XGBoost `max_depth` sensitivity (8-gen LOGO). 3-class F1 varies by only 2.0 pp; binary F1 (gen-1 vs. gen-2+) is effectively invariant ( $\leq 0.1$  pp).

| <code>max_depth</code> | 3-class F1 | Binary F1 |
|------------------------|------------|-----------|
| 3                      | 0.752      | 0.700     |
| 4                      | 0.759      | 0.699     |
| 6                      | 0.769      | 0.700     |
| 8                      | 0.772      | 0.699     |

3-class F1 increases monotonically with tree depth but the total range is only 2.0 pp (0.752–0.772); binary F1 is effectively invariant (0.699–0.700). The default `max_depth`=6 used throughout the paper lies within 0.3 pp of the best 3-class setting; the reported results are not sensitive to this choice.

### Sampling Sensitivity Analysis

To assess robustness to decoding temperature, we regenerate continuation chains at  $t \in \{0.5, 0.9\}$  with fixed nucleus  $p=0.9$  (matching the default) for four generators spanning three model families: Pythia-2.8B, OLMo-2-1B, Qwen3-1.7B, and Qwen3-0.6B (Table 14).

Table 14: Depth classification under varying temperature (fixed top- $p=0.9$ ). Binary: gen-1 vs. gen-2+. LOGO evaluation; 95% bootstrap CIs.  $^{*\dagger}$

| Model                    | Temp | LOGO Binary F1 | 95% CI       | $\rho(\text{ent. mean})$ |
|--------------------------|------|----------------|--------------|--------------------------|
| Pythia-2.8B $^{\dagger}$ | 0.5  | <b>0.814</b>   | [.798, .828] | −0.58                    |
| Pythia-2.8B $^{\dagger}$ | 0.7  | 0.740          | [.725, .755] | −0.60                    |
| Pythia-2.8B $^{\dagger}$ | 0.9  | 0.480          | [.465, .498] | −0.30                    |
| OLMo-2-1B                | 0.5  | <b>0.783</b>   | [.768, .798] | −0.72                    |
| OLMo-2-1B                | 0.7  | 0.695          | [.679, .711] | −0.60                    |
| OLMo-2-1B                | 0.9  | 0.453          | [.437, .469] | −0.26                    |
| Qwen3-1.7B $^{\dagger}$  | 0.5  | 0.751          | —            | −0.60                    |
| Qwen3-1.7B $^{\dagger}$  | 0.7  | 0.751          | —            | −0.60                    |
| Qwen3-1.7B $^{\dagger}$  | 0.9  | 0.566          | —            | −0.37                    |
| Qwen3-0.6B               | 0.5  | 0.711          | [.694, .729] | −0.67                    |
| Qwen3-0.6B               | 0.7  | 0.691          | [.673, .709] | −0.60                    |
| Qwen3-0.6B               | 0.9  | 0.598          | [.582, .615] | −0.29                    |

\*OLMo temperature experiments use OLMo-2-0425-1B; the original OLMo-2-1B-hf weights are no longer publicly available.

$^{\dagger}$ Pythia-2.8B and Qwen3-1.7B use 6-generator LOGO (pre-expansion); OLMo-2-1B and Qwen3-0.6B use 8-generator LOGO. Within-model comparisons are unaffected.

Temperature sensitivity is asymmetric and model-dependent. High-temperature degradation ( $t=0.9$ ) is universal: all four generators lose

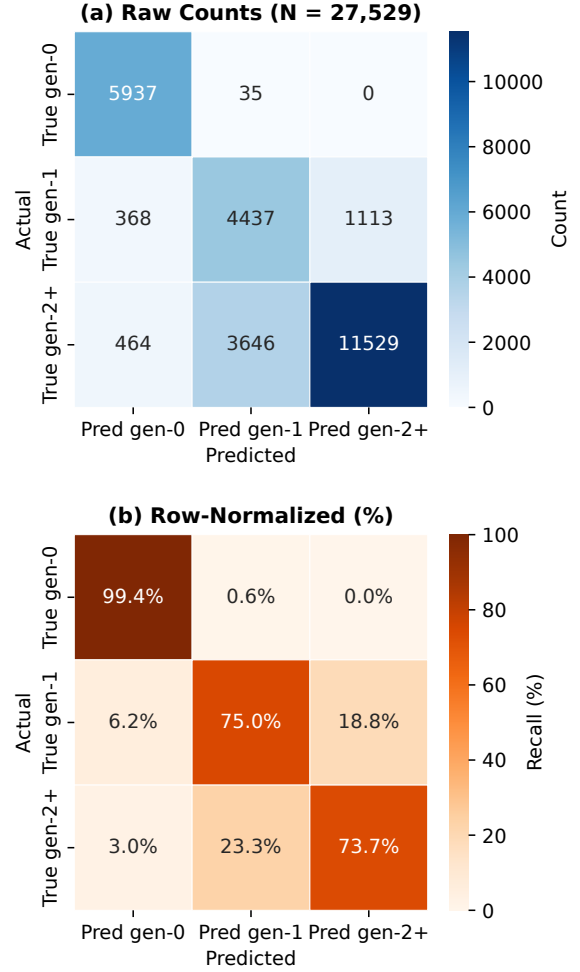


Figure 8: 3-class LOGO confusion matrix ( $N = 27,529$ ). (a) Raw counts. (b) Row-normalized recall percentages. Gen-0 achieves near-perfect classification; the primary confusion axis is gen-1  $\leftrightarrow$  gen-2+.

9.3–26.0 pp in binary F1, with entropy mean’s  $|\rho|$  weakening to 0.26–0.37 (vs. 0.58–0.72 at  $t=0.5$ ). Low-temperature improvement ( $t=0.5$ ) is model-dependent: Pythia-2.8B and OLMo-2-1B gain +7.4 and +8.8 pp respectively, while both Qwen models show negligible change (0–2.0 pp). This split tracks model family rather than scale: Qwen3-0.6B (smallest) and Qwen3-1.7B both plateau, while OLMo-2-1B (similar scale to Qwen3-1.7B) benefits substantially. One possible explanation is that Qwen models produce more peaked output distributions at  $t=0.7$ , leaving less room for low-temperature compression to sharpen the depth signal.

The method operates reliably within  $t \in [0.5, 0.7]$  across all tested models. Degradation onset at  $t=0.9$  defines the practical upper bound, consistent with the hypothesis that elevated

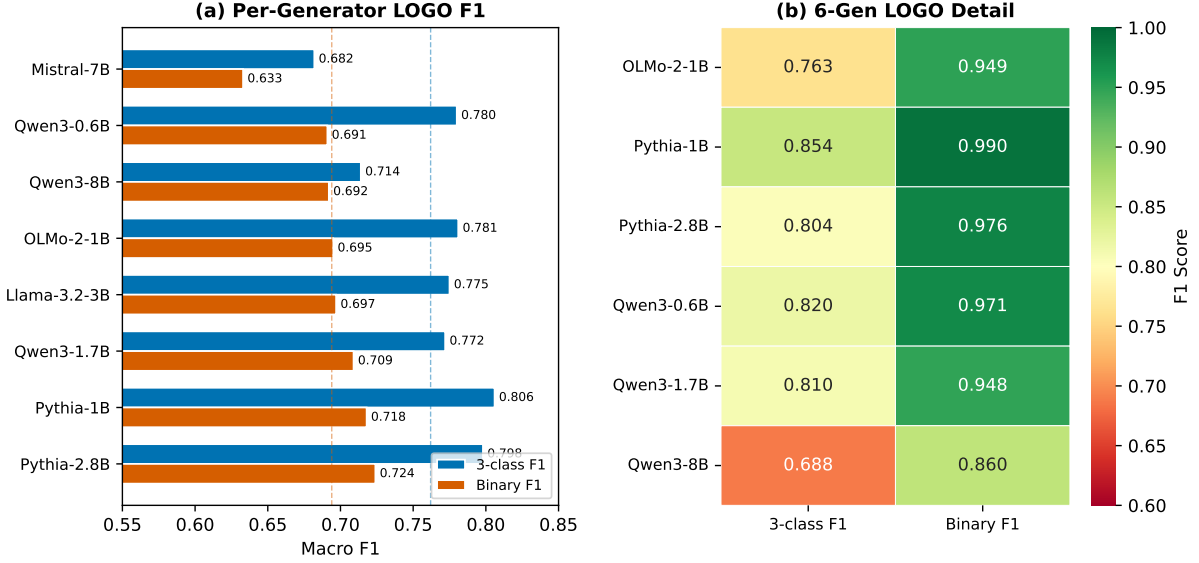


Figure 9: Per-generator LOGO performance. (a) Binary and 3-class F1, sorted by binary F1. Dashed lines: pool average. (b) 6-generator LOGO detail heatmap across model families.

temperature injects sufficient randomness to flatten the distributional drift that DDR exploits (§5).

### Per-Generator Performance

Figure 9 visualizes per-generator LOGO performance. Pythia models (1B, 2.8B) achieve highest F1, while larger models (Mistral-7B, Qwen3-8B) show degradation. The scale effect (larger models  $\rightarrow$  lower F1) is consistent with higher-capacity models generating text statistically closer to human writing.

### i Robustness to Text Perturbation

We assess robustness to post-generation text modification by randomly replacing words at varying rates and re-evaluating the 8-generator LOGO pipeline (Table 15). Random word replacement represents a worst-case perturbation: real human editing preserves semantic coherence, so practical degradation is expected to be substantially lower than these estimates.

Table 15: Robustness to random word replacement under 8-gen LOGO evaluation (binary: gen-1 vs. gen-2+). Bootstrap 95% CIs on binary F1 ( $n = 1000$ ).

| Perturbation rate | Binary F1 | 95% CI         | 3-class F1 |
|-------------------|-----------|----------------|------------|
| 0% (baseline)     | 0.695     | [0.680, 0.710] | 0.737      |
| 5%                | 0.650     | [0.635, 0.666] | 0.641      |
| 10%               | 0.614     | [0.599, 0.629] | 0.492      |
| 20%               | 0.591     | [0.576, 0.605] | 0.367      |

Binary depth discrimination (gen-1 vs. gen-2+)

degrades sublinearly: the first 5% replacement causes the largest drop ( $-4.6$  pp), with diminishing marginal impact thereafter ( $-3.6$  pp for  $5\% \rightarrow 10\%$ ,  $-2.3$  pp for  $10\% \rightarrow 20\%$ ). At 10% replacement, binary F1 remains 0.614, well above the 0.50 random baseline, which indicates that depth signal survives moderate perturbation.

Three-class classification is more sensitive ( $0.737 \rightarrow 0.367$  at 20%), because distinguishing gen-0 from gen-1 relies heavily on lexical diversity features (TTR, hapax ratio) that are directly disrupted by word replacement. The binary metric, which excludes gen-0, is inherently more robust since it depends primarily on reference-model entropy features distributed across many tokens.

**Feature subset degradation.** Table 16 compares perturbation robustness of the full 7-feature ensemble against an entropy-only subset (5 reference-model features). The higher unperturbed binary F1 of the entropy-only subset (0.833 vs. 0.695) reflects the strong discriminability of reference-model features for the coarse binary task (gen-1 vs. gen-2+); however, lexical features are essential for the finer-grained ordinal task (3-class F1: 0.737 vs. 0.688) and provide greater robustness under perturbation. At 20% replacement, the 7-feature ensemble surpasses entropy-only on binary F1 (0.591 vs. 0.566), despite starting from a lower baseline. Random word replacement drastically alters per-token context and makes entropy features (computed over the full token sequence) extremely sensitive to noise.

Lexical features (TTR, hapax ratio), as document-level summary statistics, are more stable under per-word perturbation and act as stabilizers in the full ensemble. The crossover confirms that degradation in the full feature set reflects genuine signal disruption rather than a lexical-feature artifact.

Table 16: Feature subset degradation under random word replacement (8-gen LOGO). Entropy-only: 5 reference-model features; 7-feature: full ensemble including lexical features.

| Perturb. | Ent-only Bin | 7-feat Bin | Ent-only 3-cl | 7-feat 3-cl |
|----------|--------------|------------|---------------|-------------|
| 0%       | 0.833        | 0.695      | 0.688         | 0.737       |
| 5%       | 0.736        | 0.650      | 0.539         | 0.641       |
| 10%      | 0.657        | 0.614      | 0.427         | 0.492       |
| 20%      | 0.566        | 0.591      | 0.314         | 0.367       |

**Baseline comparison.** Table 17 compares robustness across detection methods under identical perturbation conditions (8-gen LOGO, binary: gen-1 vs. gen-2+).

Table 17: Baseline robustness under random word replacement (8-gen LOGO, binary F1). Bold indicates highest F1 at each perturbation rate. (Baseline scores use default XGBoost settings; trends are consistent with the unified configuration in Table 3.)

| Method                  | 0%           | 5%           | 10%          | 20%          |
|-------------------------|--------------|--------------|--------------|--------------|
| 7-feature ensemble      | <b>0.695</b> | 0.650        | 0.614        | 0.591        |
| LD-Score                | 0.677        | <b>0.673</b> | <b>0.677</b> | <b>0.671</b> |
| Avg-LogProb             | 0.682        | 0.419        | 0.335        | 0.375        |
| FastDetectGPT           | 0.571        | 0.437        | 0.468        | 0.514        |
| Binoculars <sup>†</sup> | 0.500        | 0.461        | 0.446        | 0.425        |

<sup>†</sup>Binoculars operates at chance level on this binary task; its degradation figures are not meaningful.

The 7-feature ensemble achieves the highest unperturbed binary F1 (0.695), but LD-Score surpasses it under perturbation due to near-zero degradation ( $-0.6$  pp across the full  $0\% \rightarrow 20\%$  range). FastDetectGPT and Avg-LogProb exhibit non-monotonic degradation at higher perturbation rates, likely because severe perturbation introduces distributional artifacts that these token-level methods detect as anomalous. Avg-LogProb is the most fragile overall ( $-34.7$  pp at 10%), while FastDetectGPT drops sharply at 5% ( $-13.4$  pp) before partially recovering. LD-Score is nearly invariant to perturbation, consistent with its reliance on corpus-level distributional divergence rather than per-token likelihoods. The 7-feature ensemble offers the highest starting performance among single-reference methods, with sublinear degradation ( $-10.4$  pp at 20%).

**LLM-simulated sentence regeneration.** To test robustness under semantically coherent editing, we use Qwen3-0.6B-Base to randomly regenerate 1–3 sentences per passage (avg 1.69 sentences,  $\sim 50\%$  edit distance) and re-evaluate the 8-gen LOGO pipeline. Binary F1 drops from 0.692 to 0.647 ( $-4.5$  pp, 6.5% relative); 3-class F1 drops from 0.739 to 0.667 ( $-7.1$  pp, 9.7% relative). The binary degradation ( $-4.5$  pp) is nearly identical to that caused by 5% random word replacement ( $-4.6$  pp), despite the LLM rephrase altering  $\sim 50\%$  of characters. This asymmetry arises because LLM-generated replacements preserve the underlying statistical signature of the generation process (entropy dynamics, depth-accumulated distributional drift), whereas random word substitution directly destroys local statistical signals. Depth profiling thus captures deep generative structure that survives semantic-preserving edits.

**Seed complexity.** Performance is stable across seed text complexity: stratifying seeds by vocabulary richness (TTR terciles) yields maximum F1 variation of 1.1 pp across groups.

## j Cross-Domain Generalization

This pilot study evaluates whether depth signal generalizes beyond Wikipedia by constructing generation chains from three additional domains: News articles, Academic abstracts, and Code documentation. For each domain, we sample 200 seed texts and generate chains of depth 3 using three generators (Pythia-1B, Qwen3-1.7B, Mistral-7B). The same 7-feature ensemble and LOGO protocol are applied.

**Important caveat:** Cross-domain generation used  $t=1.0$ ,  $\text{top-}p=1.0$ , which falls outside the method’s reliable operating range ( $t \leq 0.7$ ; Appendix h). Domain and temperature effects are therefore confounded in the results below; the cross-domain signal may partly reflect temperature-induced distributional divergence rather than domain generalization alone.

Table 18: Cross-domain depth classification under LOGO evaluation (binary: gen-1 vs. gen-2+). Three generators  $\times$  200 seeds per domain. Bootstrap 95% CIs ( $n = 1000$ ).

| Domain   | Binary F1 | 95% CI         | 3-class F1 |
|----------|-----------|----------------|------------|
| Code     | 0.754     | [0.737, 0.770] | 0.634      |
| News     | 0.746     | [0.726, 0.764] | 0.652      |
| Academic | 0.739     | [0.720, 0.757] | 0.622      |

All three non-Wikipedia domains achieve binary  $F1 > 0.73$  under leave-one-generator-out evaluation. Depth signal is not an artifact of Wikipedia text properties. The Code-domain 3-class  $F1$  (0.634) likely reflects domain mismatch in the Wikipedia-trained reference model (Pythia-410M), whose NLL estimates may be less discriminative for code syntax; these cross-domain results are preliminary.

**Domain  $\times$  temperature interaction.** Cross-domain generation used  $t=1.0$ ,  $\text{top-}p=1.0$  (vs. the main experiment’s  $t=0.7$ ,  $\text{top-}p=0.9$ ). Supplementary evaluation at  $t=0.7$  reveals a domain  $\times$  temperature interaction: the main Wikipedia experiment favors lower temperature (depth signal degrades monotonically with rising temperature; Appendix h), while cross-domain evaluation favors higher temperature ( $t=1.0$  yields +3 to +11 pp over  $t=0.7$ ; News 0.746 vs. 0.712, Academic 0.739 vs. 0.628, Code 0.754 vs. 0.687). This asymmetry likely reflects reference-model calibration: Pythia-410M, trained predominantly on web text, detects subtle per-step drift in Wikipedia but requires larger distributional divergence (produced by higher temperature) to detect depth signal in out-of-distribution domains. This finding suggests that reference model selection is a key factor for cross-domain deployment of depth profiling.

**3-class inversion as DDR evidence.** In the cross-domain setting, 3-class  $F1$  (0.62–0.65) falls below binary  $F1$  (0.74–0.75), reversing the pattern observed in the main experiment (3-class 0.762 > binary 0.694). This inversion reflects degraded gen-0 vs. gen-1 separation: the Pythia-410M reference model is calibrated to Wikipedia text, so its baseline entropy estimates are less discriminative for out-of-domain human text and compresses the feature distance between gen-0 and gen-1. The binary task (gen-1 vs. gen-2+) is unaffected because both classes are generated under identical domain and decoding conditions; recursive drift is a relative change that does not depend on absolute calibration. This pattern provides additional evidence that DDR captures base-model generation dynamics rather than domain-specific text properties: depth signal remains robust even when the human-vs-machine boundary becomes less distinct.

## k Instruction-Tuned Model Analysis

### Instruction-Tuned Model Results

Table 19 summarizes depth classification under instruction tuning for three model families. All results use within-model 5-fold CV on single-generator chains; LOGO evaluation is available for two models but uses earlier generator subsets (5–6 generators), so cross-model numbers are not directly comparable across rows.

Table 19: Depth classification: base vs. instruction-tuned models (within-model 5-fold CV).  $\rho$ : Spearman correlation of entropy mean with depth. Rev.: number of features (out of 7) whose  $\rho$  reverses sign under instruction tuning.

| Model        | 3-class F1 |          | Binary F1 |          | $\rho(\text{ent. mean})$ |          | Rev. |
|--------------|------------|----------|-----------|----------|--------------------------|----------|------|
|              | Base       | Instruct | Base      | Instruct | Base                     | Instruct |      |
| OLMo-2-1B    | 0.754      | 0.575    | 0.845     | 0.813    | −0.64                    | +0.48    | 6/7  |
| Qwen3-1.7B   | 0.769      | 0.732    | 0.861     | 0.875    | −0.60                    | +0.48    | 6/7  |
| Llama-3.2-3B | 0.742      | 0.544    | 0.849     | 0.838    | −0.63                    | +0.23    | 3/7  |

Instruction tuning degrades 3-class  $F1$  by 3.7–19.9 pp across all three families, while binary  $F1$  (gen-1 vs. gen-2+) changes remain small ( $|\Delta| \leq 3.2$  pp). The asymmetry indicates that instruction tuning primarily disrupts the gen-0/gen-1 boundary (human vs. first-generation), consistent with RLHF shifting the output distribution away from human-like text. Entropy-mean  $\rho$  reverses sign for all three models (−0.60–−0.64 base  $\rightarrow$  +0.23–+0.48 instruct), and 6 of 7 features reverse for OLMo and Qwen vs. 3 of 7 for Llama. This pattern is consistent with instruction tuning shifting the generation attractor rather than eliminating iterative dynamics (§5).

**Retraining on instruction-tuned data.** To test whether the depth signal persists under instruction tuning with reversed polarity, we retrain the XGBoost classifier on instruction-tuned feature data (Table 20).

Retraining recovers substantial within-model performance (+6.8–26.9 pp), but the retrained  $F1$  (0.590–0.660) remains below base-model levels (0.69–0.72). Instruction tuning compresses depth signal beyond what retraining on reversed-polarity features can fully recover.

**DDR structure under instruction tuning.**  $k=2$  clustering in  $(|\rho|, d)$  space on pooled instruct features yields silhouette = 0.372 (vs. 0.660 for base models); the clean drifting/non-drifting partition is substantially compressed. The compression

Table 20: Classification recovery under instruction-tuned retraining (binary-only: gen-1 vs. gen-2+, gen-0 excluded from training and evaluation). Base→Inst: base-trained classifier tested on instruct data. Note: Table 19 reports higher instruct binary F1 (e.g., OLMo 0.813) because it uses 3-class training with binary evaluation, where gen-0 in training helps the classifier learn the feature space.

| Model           | Base→Inst | Retrained | $\Delta$ |
|-----------------|-----------|-----------|----------|
| OLMo-2-1B-Inst  | 0.321     | 0.590     | +0.269   |
| Qwen3-1.7B-Inst | 0.500     | 0.660     | +0.160   |
| Qwen3-0.6B-Inst | 0.532     | 0.600     | +0.068   |

is model-dependent: OLMo-Instruct shows near-zero  $|\rho|$  for TTR and hapax (0.006, 0.025), while Qwen instruct models retain moderate correlations ( $|\rho|=0.48\text{--}0.53$ ). Cross-condition transfer is near chance in both directions (base→instruct binary F1 = 0.490; instruct→base F1 = 0.454), with the asymmetry ( $-0.036$ ) suggesting that instruct-trained features generalize even less to base distributions than the reverse. These results indicate that instruction tuning compresses the feature-space structure underlying DDR rather than simply reversing feature directions (Figure 10).

**Reference model sensitivity.** One hypothesis for the instruct degradation is reference-model mismatch: the base Pythia-410M reference may produce systematically biased NLL estimates on instruction-tuned outputs, masking depth-dependent variation. To test this, we replace the reference with Qwen3-0.6B-Instruct and recompute all features on the two instruct generators. The hypothesis is partially rejected: Qwen3-1.7B-Instruct shows mixed changes (entropy max  $d$ :  $0.51 \rightarrow 0.08$ , becoming near-stationary; entropy std  $d$ :  $1.07 \rightarrow 1.51$ , increasing), while OLMo-Instruct loses nearly all depth signal (entropy mean  $d$ :  $1.33 \rightarrow 0.02$ ; tail decay  $d$ :  $0.60 \rightarrow 0.06$ ). The model-dependent heterogeneity rules out reference mismatch as a universal explanation for instruct DDR degradation. Alignment fundamentally reshapes the generation dynamics rather than introducing a correctable measurement artifact.

**End-to-end instruct classification.** To test whether depth signal persists at the token level despite summary-statistic degradation, we apply the QLoRA end-to-end classifier (Appendix e) to instruction-tuned data using within-model 5-fold stratified CV (Table 21).

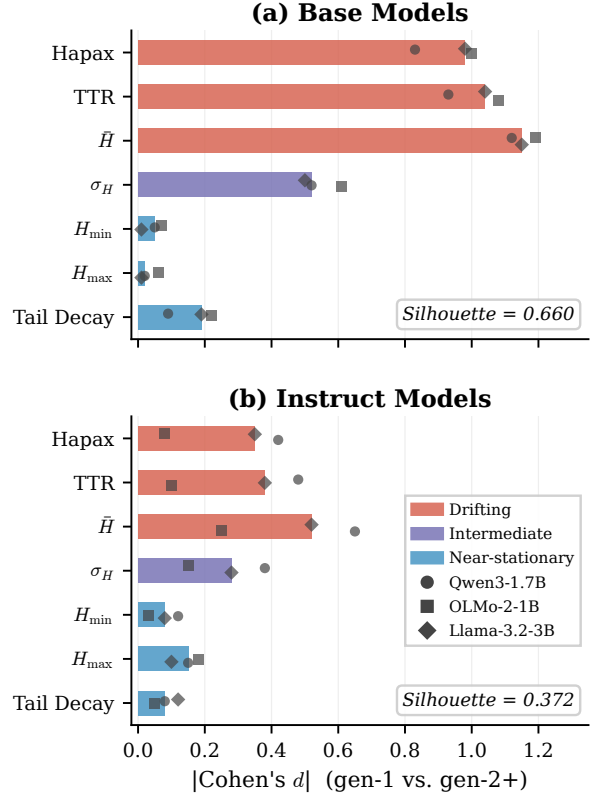


Figure 10: DDR partition under base vs. instruction-tuned models. (a) Base models (silhouette = 0.660): clear drifting/non-drifting separation. (b) Instruction-tuned models (silhouette = 0.372): all features compress toward zero, collapsing the partition; entropy max shifts from near-stationary to intermediate; formerly drifting features (hapax, TTR) lose most of their effect size. Bars show pooled  $|d|$ ; markers show individual generators.

The XGBoost-to-QLoRA gap on the same instruct data is substantial (+17.5 to +31.3 pp). Depth signal persists at the token level under instruction tuning but is inaccessible to the seven summary statistics; the degradation reported in Table 19 reflects a feature-level limitation rather than signal absence. Note that instruct samples are approximately twice the base count used here (Table 21), so the instruct-vs-base QLoRA comparison is confounded by training set size; the meaningful comparison is QLoRA-vs-XGBoost on the same instruct data. All results use within-model 5-fold CV, not cross-generator LOGO, limiting generalization claims.

## 1 Error Analysis of Binary Classification

We analyze misclassification patterns in the binary LOGO task (gen-1 vs. gen-2+) across 29,095 test samples (overall error rate 28.1%).

Table 21: End-to-end QLoRA vs. XGBoost on instruction-tuned data (within-model 5-fold CV, Qwen3-0.6B backbone). Base QLoRA rows shown for reference only; instruct sets ( $N \approx 4.4\text{--}5.0\text{K}$ ) are approximately twice the base sets ( $N \approx 2.3\text{K}$ ), so cross-condition comparison is confounded by training set size.

| Model             | 3-class F1 |                   | $\Delta$ | $N$   |
|-------------------|------------|-------------------|----------|-------|
|                   | XGB        | QLoRA             |          |       |
| OLMo-2-1B-Inst    | 0.575      | $0.888 \pm 0.007$ | +31.3 pp | 4,429 |
| Qwen3-1.7B-Inst   | 0.732      | $0.907 \pm 0.008$ | +17.5 pp | 4,963 |
| OLMo-2-1B (base)  |            | $0.859 \pm 0.004$ |          | 2,379 |
| Qwen3-1.7B (base) |            | $0.904 \pm 0.009$ |          | 2,291 |

**Error asymmetry.** The error distribution is highly asymmetric: 86.2% of errors are false negatives (gen-2+ predicted as gen-1), while only 13.8% are false positives (Table 22). The classifier systematically under-predicts depth, consistent with the DDR pattern: shallow gen-2+ samples that have not yet accumulated sufficient distributional drift are indistinguishable from gen-1. Among generators, Mistral-7B shows the highest error rate (45.0%), while Pythia-1B is easiest (18.2%). Larger models produce gen-2+ text closer to the gen-1 distribution.

Table 22: Binary classification error analysis (gen-1 vs. gen-2+) under LOGO cross-validation. FP: gen-1 misclassified as gen-2+; FN: gen-2+ misclassified as gen-1. Errors are dominated by false negatives (86.2%).

| Generator      | $N$   | FP   | FN   | Err.% |
|----------------|-------|------|------|-------|
| Pythia-1B      | 3560  | 188  | 460  | 18.2  |
| Pythia-2.8B    | 3614  | 129  | 737  | 24.0  |
| OLMo-2-1B      | 3718  | 154  | 900  | 28.3  |
| Qwen3-0.6B     | 3542  | 234  | 511  | 21.0  |
| Qwen3-1.7B     | 3621  | 175  | 639  | 22.5  |
| Qwen3-8B       | 3502  | 93   | 1104 | 34.2  |
| Llama-3.2-3B   | 3685  | 82   | 1030 | 30.2  |
| Mistral-7B     | 3853  | 75   | 1659 | 45.0  |
| <b>Overall</b> | 29095 | 1130 | 7040 | 28.1  |

**Per-depth error rates.** Error rates decrease monotonically with depth (Table 23): depth-2 samples are the primary failure point (48.4% error rate), followed by depth-3 (29.2%) and depth-4 (20.3%). The monotonic decrease confirms that deeper generations accumulate progressively stronger depth signatures; the gen-1  $\rightarrow$  gen-2 transition is the hardest boundary.

**Feature-level explanation.** Misclassified samples occupy the overlap zone between gen-1 and gen-2+ distributions (Table 24): they exhibit ele-

Table 23: Error rate by generation depth in binary classification. Depth-2 exhibits the highest misclassification rate (48.4%); the gen-1  $\rightarrow$  gen-2 transition is the primary confusion boundary.

| Depth | $N$  | Errors | Rate (%) | Type |
|-------|------|--------|----------|------|
| 1     | 7896 | 1130   | 14.3     | FP   |
| 2     | 7512 | 3637   | 48.4     | FN   |
| 3     | 7032 | 2051   | 29.2     | FN   |
| 4     | 6655 | 1352   | 20.3     | FN   |

vated entropy ( $\bar{H}$ : 2.26 vs. 1.57,  $p < 10^{-300}$ ) and lexical diversity (TTR: 0.464 vs. 0.302), retaining gen-1-like characteristics. In contrast, shape features show no discriminating power: entropy max ( $p = \text{n.s.}$ ) and tail decay ( $p = \text{n.s.}$ ), consistent with the DDR pattern (§4.2).

Table 24: Feature comparison between correctly and incorrectly classified samples. Misclassified samples show substantially higher entropy and lexical diversity but no difference in shape features.

| Feature               | Correct | Incorrect | $\Delta$ | $p$                   |
|-----------------------|---------|-----------|----------|-----------------------|
| $\bar{H}$             | 1.5700  | 2.2618    | +0.6918  | $< 10^{-300} ***$     |
| $\sigma_H$            | 1.6227  | 1.7413    | +0.1186  | $< 10^{-300} ***$     |
| $H_{\min}$            | 0.0281  | 0.0330    | +0.0049  | $1.0\text{e-}294 ***$ |
| $H_{\max}$            | 7.5255  | 7.5251    | -0.0004  | n.s.                  |
| $\beta_{\text{tail}}$ | -0.0023 | -0.0023   | +0.0000  | n.s.                  |
| TTR                   | 0.3024  | 0.4635    | +0.1612  | $< 10^{-300} ***$     |
| Hapax                 | 0.1760  | 0.3131    | +0.1371  | $< 10^{-300} ***$     |

**3-class confusion matrix.** Table 25 shows the aggregated confusion matrix: gen-0 is near-perfectly classified ( $>99\%$ ), while gen-1 and gen-2+ exhibit substantial cross-contamination (30.2% of gen-2+ predicted as gen-1), consistent with the binary error analysis.

Table 25: Aggregated 3-class confusion matrix (%) across 8-fold LOGO. Gen-0 is near-perfectly classified, while gen-2+ is frequently confused with gen-1.

| True   | Predicted |       |        |
|--------|-----------|-------|--------|
|        | gen-0     | gen-1 | gen-2+ |
| gen-0  | 99.2      | 0.8   | 0.1    |
| gen-1  | 7.3       | 78.4  | 14.3   |
| gen-2+ | 3.0       | 30.2  | 66.7   |

## m Downstream Utility

Can depth-aware data curation improve downstream language model quality? We fine-tune OLMo-2-0425-1B on five data variants (3 seeds

each) and evaluate perplexity on a held-out test set (Table 26).

Table 26: Downstream fine-tuning perplexity (OLMo-2-0425-1B, mean  $\pm$  std across 3 seeds). At equal sample count, random subsampling (2.367) outperforms depth-aware curation (depth 0–1: 2.387, binary-filtered: 2.389).

| Variant               | $n$    | Test PPL          |
|-----------------------|--------|-------------------|
| All depths            | 12,299 | 2.335 $\pm$ 0.000 |
| Size-matched (random) | 3,951  | 2.367 $\pm$ 0.000 |
| Depth 0–1 only        | 3,951  | 2.387 $\pm$ 0.001 |
| Binary-filtered       | 3,951  | 2.389 $\pm$ 0.000 |
| Depth 0 only          | 797    | 2.681 $\pm$ 0.001 |

The results are unambiguous: data quantity dominates depth composition. Training on all 12,299 samples yields the lowest perplexity (2.335). At equal sample count ( $n = 3,951$ ), random subsampling (2.367) *outperforms* both depth-stratified selection (depth 0–1: 2.387) and binary detection filtering (2.389). Cross-seed variance is below 0.002 across all variants, which confirms the robustness of this null result.

Restricting training to human-only text (depth 0,  $n = 797$ ) produces a sharp perplexity increase to 2.681, driven primarily by the  $5\times$  reduction in data volume rather than depth composition per se.

**Equivalence testing.** We apply a two one-sided  $t$ -test (TOST) with a  $\pm 2\%$  relative margin (equivalence bounds  $\pm 0.047$  on the ppl scale) to formally assess whether the depth-aware and random subsampling conditions produce practically indistinguishable perplexity. At equal sample count ( $n = 3,951$ ): size-matched (random) ppl = 2.367 vs. depth 0–1 ppl = 2.387,  $\Delta = 0.020$ . TOST confirms equivalence ( $p = 1.1 \times 10^{-4}$ ); the two curation strategies yield practically indistinguishable downstream quality within the  $\pm 2\%$  margin. Cross-seed standard deviation is below 0.001 for both variants.

Comparing the full dataset ( $n = 12,299$ , ppl = 2.335) against the depth-filtered subset ( $n = 3,951$ , ppl = 2.387): TOST *rejects* equivalence (minimum equivalence margin = 0.053, exceeding the  $\pm 2\%$  bound), which confirms that data quantity drives perplexity reduction beyond what depth-aware selection can recover.

Knowing the generation depth of a document does not, by itself, make it more or less useful for training; at fixed sample count, depth-aware curation is statistically equivalent to uninformed random subsampling.

## Margin Sensitivity Analysis

To assess the robustness of the equivalence conclusion to margin choice, we repeat the TOST procedure across multiple equivalence margins (Lakens, 2017), following the recommendation to report sensitivity across plausible bounds rather than relying on a single threshold (Table 27).

Table 27: TOST equivalence testing across multiple margins. “Equiv?” indicates whether equivalence is established at  $\alpha = 0.05$ .

| Comparison                          | Margin    | Equiv? | TOST $p$           |
|-------------------------------------|-----------|--------|--------------------|
| Full (12,299) vs. Random (3,951)    | $\pm 1\%$ | No     | 1.0                |
|                                     | $\pm 2\%$ | Yes    | $1 \times 10^{-6}$ |
|                                     | $\pm 3\%$ | Yes    | $< 10^{-10}$       |
|                                     | $\pm 5\%$ | Yes    | $< 10^{-10}$       |
| Full (12,299) vs. Depth 0–1 (3,951) | $\pm 1\%$ | No     | 1.0                |
|                                     | $\pm 2\%$ | No     | 1.0                |
|                                     | $\pm 3\%$ | Yes    | $< 10^{-10}$       |
|                                     | $\pm 5\%$ | Yes    | $< 10^{-10}$       |

The full dataset vs. random subsample comparison passes equivalence at the  $\pm 2\%$  margin ( $p = 10^{-6}$ ), while the full vs. depth 0–1 comparison requires a  $\pm 3\%$  margin. In absolute terms, perplexity differences are small:  $\Delta = 0.032$  (full vs. random, 1.4% of baseline) and  $\Delta = 0.052$  (full vs. depth 0–1, 2.2% of baseline).

Cohen’s  $d$  values are 86.0 (full vs. random) and 119.7 (full vs. depth 0–1). These appear large but are unreliable: with only  $n = 3$  seeds per condition, the pooled standard deviation is extremely small ( $< 0.001$ ), artificially inflating  $d$ . The inflated effect sizes reflect low seed-level variance (confirming reproducibility) rather than practically meaningful perplexity differences. The TOST framework with pre-specified relative margins provides a more appropriate assessment than Cohen’s  $d$  in this low-seed regime. However, with only  $n = 3$  seeds per condition, the TOST results have limited generalizability: the test is valid given observed variance but the conclusion generalizes over only three random seeds. We characterize this as a preliminary observation rather than a definitive null result.

## n Cross-Model Chain Per-Class Detail

This appendix reports per-generator breakdowns for the cross-model chain evaluation described in §4.1.

**Variance decomposition.** A two-way ANOVA (generator  $\times$  depth) on cross-model chain features ( $n = 3,778$ , depths 1–4) confirms that generator identity contributes negligible variance: partial

Table 28: Cross-model chain per-class performance (generator-level leave-one-out,  $N=2,964$ ,  $n_{\text{generators}}=5$ : Pythia-2.8B/1B, OLMo-2-1B, Qwen3-0.6B/1.7B). Macro F1=0.831 under class balancing; under natural class distribution, macro F1=0.789. Binary F1 (gen-1 vs. gen-2+)=0.830 (95% CI [.818, .841], case-level bootstrap,  $n=2,000$ ). Cross-model chains avoid the gen-0 seed sharing issue present in single-generator LOGO.

| Class        | Prec. | Rec.  | F1           | 95% CI       |
|--------------|-------|-------|--------------|--------------|
| gen-0        | 0.921 | 0.948 | 0.935        | [.923, .946] |
| gen-1        | 0.737 | 0.803 | 0.768        | [.748, .788] |
| gen-2+       | 0.842 | 0.742 | 0.789        | [.768, .809] |
| <b>Macro</b> | —     | —     | <b>0.831</b> | [.817, .844] |

$\eta_{\text{gen}}^2 < 1\%$  for all seven features (range 0.05%–0.90%, mean 0.37%), while depth accounts for 15.3%–29.9% of variance in drifting features. The generator $\times$ depth interaction is also  $< 1\%$ . This confirms that the F1 = 0.830 cross-model advantage reflects depth signal rather than generator diversity accumulation: mixing generators across chain steps does not introduce a confounding generator-identity cue.

## o Feature Distribution Analysis

### DDR Trend Lines

Figure 11 extends the main-text DDR trajectory plot (Figure 1) by showing per-feature depth trends with individual generator traces overlaid.  $R^2$  values are computed from ordinary least-squares regression on depth-level means (five points per generator, one per depth 0–4), not per-sample regression; the annotated Spearman  $\rho$  provides the complementary per-sample monotonicity measure.

### UMAP Embedding

Figure 12 shows a 2D UMAP projection of the 7-feature space (7,500 stratified samples, StandardScaler normalization,  $k=30$  neighbors, min\_dist=0.3). Panel (a) reveals a depth gradient from gen-0 through gen-4, with progressive overlap at deeper generations consistent with the convergence that motivates the gen-2+ class collapse. Panel (b) shows the 3-class view: gen-0 forms a separable cluster, while gen-1 and gen-2+ overlap substantially, explaining the classification difficulty at this boundary.

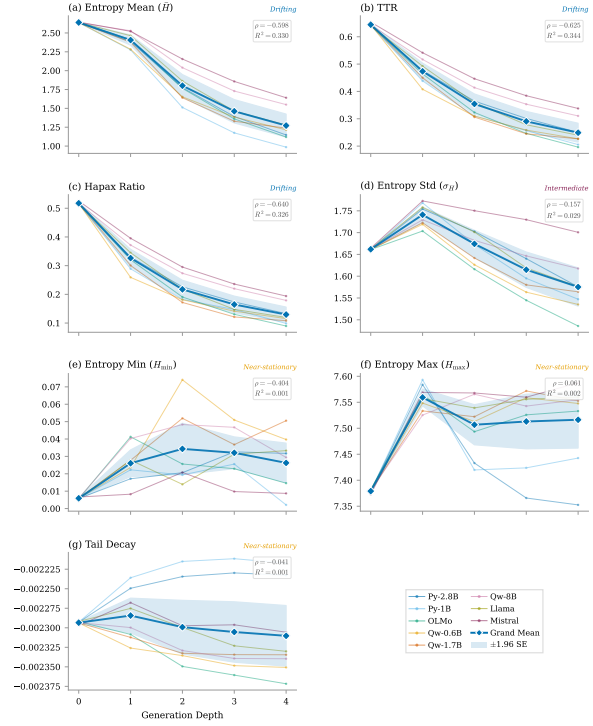


Figure 11: Per-feature depth trends. Bold line: pooled mean ( $\pm 1.96$  SE shading) across all generators, colored by DDR zone (blue: drifting, wine: intermediate, amber: near-stationary). Thin traces: individual generators. Spearman  $\rho$  and linear  $R^2$  are annotated per panel.

## p Cross-Temperature Transfer

Figure 13 shows cross-temperature transfer results for the 6-generator LOGO pipeline. We evaluate four train $\rightarrow$ test temperature conditions using Qwen3-1.7B data regenerated at  $t=1.0$  (remaining five generators retain  $t=0.7$ ).<sup>6</sup>

Key observations from the cross-temperature analysis:

- Matched-temperature robustness:** Both  $t=0.7$  and  $t=1.0$  matched conditions yield comparable performance (3-class F1: 0.789 vs. 0.726), with the gap attributable to weaker depth signal at high temperature rather than classifier degradation.
- Asymmetric transfer:** Training at  $t=1.0$  and testing at  $t=0.7$  (F1=0.791) transfers well, but the reverse ( $0.7\rightarrow 1.0$ , F1=0.723) degrades. This asymmetry reflects the information structure: features learned from high-variance (high-temperature) data generalize to low-variance settings, but not vice versa.

<sup>6</sup>The  $t=1.0$  data uses top- $p=0.95$  (vs.  $p=0.9$  for other configurations); this difference may contribute to the observed degradation beyond the temperature effect alone.

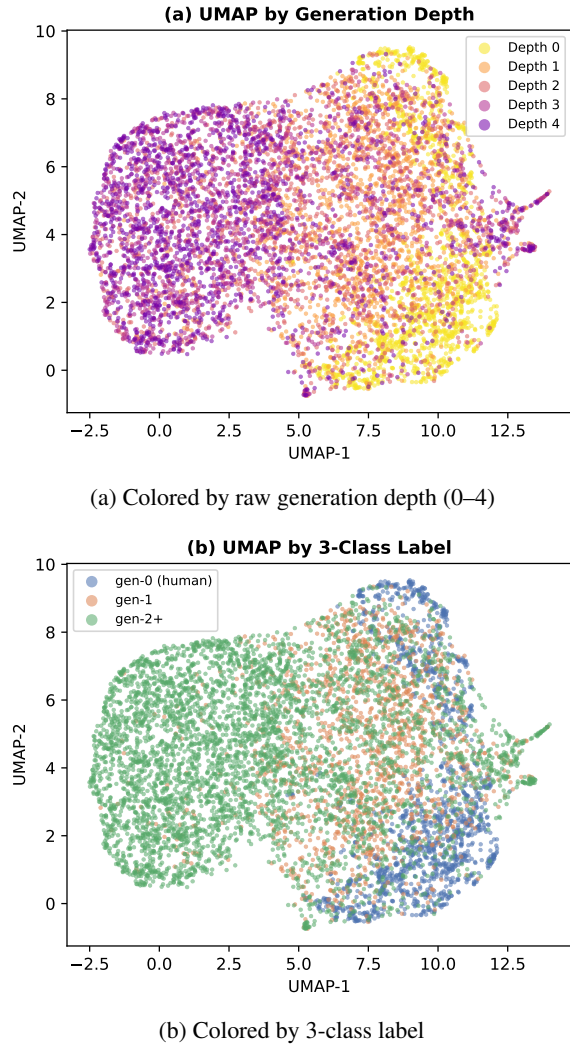


Figure 12: UMAP 2D projection of the 7-feature space ( $N = 7,500$  stratified sample). (a) A gradient structure emerges, with deeper depths converging. (b) Gen-0 separates clearly; gen-1 and gen-2+ overlap, consistent with the high false-negative rate at the gen-1/gen-2+ boundary.

- Generator-specific vulnerability:** Qwen3-1.7B, the only generator with temperature-mismatched test data, drops to  $F1 = 0.418$  under cross-temperature conditions while all other generators maintain  $F1 > 0.76$ . This confirms that depth features are calibrated to the specific decoding regime.
- Binary transfer for the mismatched generator:** For Qwen3-1.7B (the only generator with actual temperature mismatch), binary F1 drops from 0.948 to 0.713 under  $0.7 \rightarrow 1.0$  transfer but recovers in the reverse direction ( $1.0 \rightarrow 0.7$ : 0.948), consistent with the asymmetric pattern in Item 2. The aggregate binary range (0.900–0.958) is inflated by five genera-

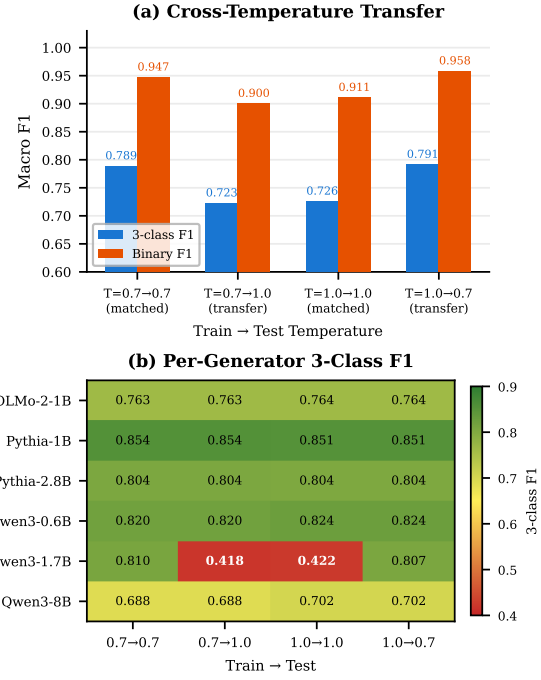


Figure 13: Cross-temperature transfer evaluation (6-gen LOGO). (a) Overall F1 by train→test temperature condition. Matched conditions ( $0.7 \rightarrow 0.7$ ,  $1.0 \rightarrow 1.0$ ) outperform cross-temperature transfer, with  $0.7 \rightarrow 1.0$  showing the largest degradation (−6.6 pp in 3-class F1). (b) Per-generator 3-class F1 heatmap: Qwen3-1.7B shows dramatic degradation under cross-temperature conditions (0.418) while other generators remain stable. The temperature mismatch disproportionately affects the generator whose data was regenerated.

tors whose data is identical across conditions.

## q Feature Ablation

This appendix presents feature ablation results supporting the complementarity analysis in §4.2.

## r Feature Correlation Matrix

This appendix documents inter-feature correlations and the PCA analysis referenced in §4.2.

TTR and hapax ratio are near-perfectly correlated ( $r = 0.97$ ), and both correlate strongly with mean entropy ( $r = 0.86$  and  $0.81$ ). These three drifting features form a tightly coupled cluster: PCA on the standardized drifting-feature subset shows that a single principal component explains 92.3% of their joint variance, with near-uniform loadings (0.56–0.59). These three features thus share a single latent dimension. Full 7-feature PCA yields PC1 (44.4% variance) loading on the drifting triplet (0.53–0.55) and PC2 (18.3% variance) capturing entropy distributional shape (entropy std

Table 29: Feature ablation (8-gen, 10-fold stratified CV, 3-class). Full > surface (+8.5 pp) and Full > reference (+6.0 pp) confirms complementarity.

| Feature set        | F1    | Top feature (imp.)  |
|--------------------|-------|---------------------|
| Full (7 features)  | 0.773 | hapax_ratio (0.35)  |
| Surface only (2)   | 0.688 | hapax_ratio (0.41)  |
| Reference only (5) | 0.713 | entropy_mean (0.27) |

Table 30: Pearson correlation matrix of the seven depth features ( $N = 27,529$ ).

|            | $\bar{H}$   | $\sigma_H$ | $H_{\min}$ | $H_{\max}$ | $\tau$ | TTR         | Hapax       |
|------------|-------------|------------|------------|------------|--------|-------------|-------------|
| $\bar{H}$  | —           | 0.49       | 0.31       | 0.01       | 0.07   | <b>0.86</b> | <b>0.81</b> |
| $\sigma_H$ | 0.49        | —          | -0.15      | 0.16       | 0.26   | 0.43        | 0.34        |
| $H_{\min}$ | 0.31        | -0.15      | —          | -0.04      | 0.01   | 0.20        | 0.23        |
| $H_{\max}$ | 0.01        | 0.16       | -0.04      | —          | -0.08  | -0.13       | -0.15       |
| $\tau$     | 0.07        | 0.26       | 0.01       | -0.08      | —      | 0.13        | 0.14        |
| TTR        | <b>0.86</b> | 0.43       | 0.20       | -0.13      | 0.13   | —           | <b>0.97</b> |
| Hapax      | <b>0.81</b> | 0.34       | 0.23       | -0.15      | 0.14   | <b>0.97</b> | —           |

vs. entropy min contrast). In contrast,  $H_{\max}$ ,  $\tau$  (tail decay), and  $H_{\min}$  show weak pairwise correlations with all other features ( $|r| < 0.27$ ); these features capture independent aspects of the token-probability distribution.

## s Cross-Family Reference Models

To isolate reference-model bias, we replace Pythia-410M with two alternative references from different families and evaluate on a 5-generator subset (Pythia-2.8B/1B, OLMo-2-1B, Qwen3-0.6B/1.7B). Table 31 and Figure 14 show that performance spread across references is only 0.013 F1, confirming the DDR partition is reference-model agnostic.

## t Independent-Seed Control

To test for topic-level leakage from shared seeds across generators, we construct an independent-seed evaluation where Pythia-2.8B and OLMo-2-1B use completely disjoint Wikipedia seed pools (500 seeds each, zero overlap). Each generator produces depth-0 through depth-4 chains on its own seed pool; features are extracted with the same Pythia-410M reference model. We then train a 7-feature XGBoost classifier on one generator’s data and test on the other, in both directions.

The small gap ( $-1.4$  pp) between independent-seed and shared-seed evaluation falls within bootstrap confidence intervals, and both directions individually exceed 0.66. Combined with the seed-level demeaning control (§4.1), these results provide converging evidence that depth classification

Table 31: Cross-family reference model validation (5-gen LOGO subset). Drift direction is consistent (negative) across all three references; magnitude varies with capacity. Spread in LOGO F1 is 0.013.

| Reference   | Family | Params | LOGO F1 | Ent. Drift       |
|-------------|--------|--------|---------|------------------|
| Pythia-410M | Pythia | 410M   | 0.803   | $[-1.19, -1.40]$ |
| OLMo-2-1B   | OLMo   | 1.5B   | 0.807   | $[-0.61, -0.91]$ |
| Qwen3-0.6B  | Qwen   | 0.6B   | 0.816   | $[-0.96, -1.35]$ |

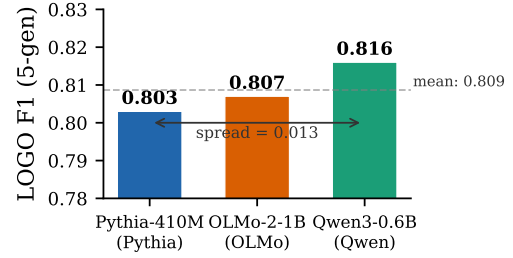


Figure 14: Cross-family reference model comparison. LOGO F1 spread is only 0.013 across three reference models from different families, confirming DDR is reference-model agnostic.

is driven by distributional properties of recursive generation, not topic identity.

**Semantic similarity analysis.** Embedding cosine similarity between consecutive depth steps increases monotonically (Figure 15):  $0 \rightarrow 1$  (0.622),  $1 \rightarrow 2$  (0.683),  $2 \rightarrow 3$  (0.801),  $3 \rightarrow 4$  (0.859). Deeper generations are thus *more* semantically similar to their inputs, yet the classifier still distinguishes depth classes. The depth signal resides in distributional rather than topical properties. Prompt type modulates the baseline similarity level (continuation  $0.717 <$  elaboration  $0.809 \approx$  rewriting  $0.808$ ), plausibly because semantic-preservation constraints reduce topic drift while leaving the distributional location shift intact.

## u Per-Generator Equivalence Testing

To confirm that the near-stationary classification is not a pooling artifact, we apply two one-sided  $t$ -tests (TOST) (Lakens, 2017) for each feature within each generator separately (gen-1 vs. gen-2+, equivalence bound  $|d| < 0.20$ ). We adopt Cohen’s conventional small effect size ( $d=0.20$ ) (Cohen, 1988) as the smallest effect of practical significance (SESOI): effects below this threshold are unlikely to be practically meaningful for depth discrimination. LOFO analysis independently supports this choice: removing any near-stationary feature changes classification F1 by  $\leq 1.5$  pp (Ap-

Table 32: Independent-seed cross-generator evaluation (binary: gen-1 vs. gen-2+). Disjoint seed pools eliminate any possibility of topic-based leakage. The average independent-seed F1 (0.680) is within 1.4 pp of the shared-seed LOGO baseline (0.694). The classifier thus captures generative degradation patterns rather than seed-specific topic content.

| Setting                             | Binary F1 | 95% CI         |
|-------------------------------------|-----------|----------------|
| Pythia-2.8B $\rightarrow$ OLMo-2-1B | 0.695     | [0.673, 0.717] |
| OLMo-2-1B $\rightarrow$ Pythia-2.8B | 0.666     | [0.643, 0.689] |
| Independent-seed average            | 0.680     | —              |
| Shared-seed LOGO (reference)        | 0.694     | [0.688, 0.700] |

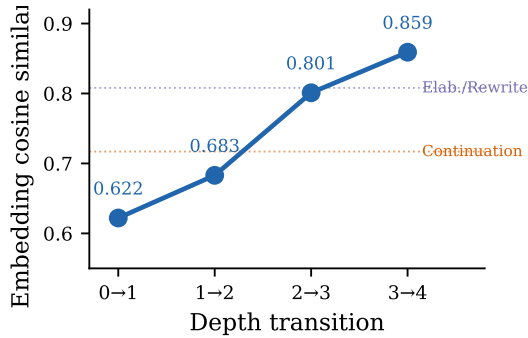


Figure 15: Embedding cosine similarity between consecutive depth steps. Deeper generations become increasingly similar to their inputs, yet the classifier still distinguishes depth classes. Dotted lines: prompt-type baselines.

pendix f), linking the statistical threshold to classification irrelevance. Sensitivity analysis at  $\varepsilon \in \{0.15, 0.20, 0.25\}$  confirms the partition is robust to the bound choice (Appendix v); bootstrap resampling ( $B=1,000$ ) shows 100% rejection stability for all near-stationary features (Appendix u.1); and the partition is invariant across  $\varepsilon \in [0.10, 0.75]$  (Appendix v). Table 33 reports observed  $|d|$ ;  $\checkmark$  marks features for which the TOST null hypothesis (non-equivalence) is rejected at  $\alpha = 0.05$ .

Entropy min passes equivalence for all eight generators; entropy max passes for six of eight (exceeding  $\varepsilon$  only for Pythia-1B and Pythia-2.8B, where  $|d| \approx 0.21$ – $0.25$ ); tail decay passes for four of eight. All drifting and intermediate features fail equivalence uniformly, with  $|d|$  ranging from 0.23 (entropy std, Mistral) to 1.40 (entropy mean, Pythia-1B). The partition is robust at the individual-generator level, with entropy min as the most consistently stationary feature. Figure 17 visualizes the DDR partition in  $(|\rho|, d)$  space; the three groups are separated by clear gaps.

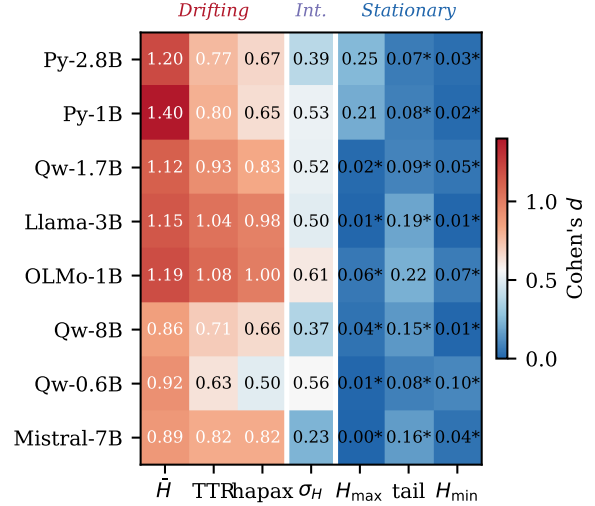


Figure 16: TOST equivalence heatmap. Cohen’s  $d$  (gen-1 vs. gen-2+) per generator and feature. Asterisks (\*) mark TOST equivalence at  $\varepsilon=0.20$ . Drifting features (red, left) show uniformly large  $d$ ; near-stationary features (blue, right) cluster near zero.

Table 33: Per-generator TOST equivalence ( $\varepsilon=|d|<0.20$ , gen-1 vs. gen-2+). Observed  $|d|$ ;  $\checkmark$  = equivalent ( $p_{\text{TOST}} < 0.05$ ). Features grouped by DDR partition.

| Generator    | Drifting  |      |       | Int.       | Near-stationary   |                   |                   |
|--------------|-----------|------|-------|------------|-------------------|-------------------|-------------------|
|              | $\bar{H}$ | TTR  | hapax | $\sigma_H$ | $H_{\max}$        | tail              | $H_{\min}$        |
| Pythia-2.8B  | 1.20      | 0.77 | 0.67  | 0.39       | 0.25              | 0.07 $\checkmark$ | 0.03 $\checkmark$ |
| Pythia-1B    | 1.40      | 0.80 | 0.65  | 0.53       | 0.21              | 0.08 $\checkmark$ | 0.02 $\checkmark$ |
| Qwen3-1.7B   | 1.12      | 0.93 | 0.83  | 0.52       | 0.02 $\checkmark$ | 0.09 $\checkmark$ | 0.05 $\checkmark$ |
| Llama-3.2-3B | 1.15      | 1.04 | 0.98  | 0.50       | 0.01 $\checkmark$ | 0.19              | 0.01 $\checkmark$ |
| OLMo-2-1B    | 1.19      | 1.08 | 1.00  | 0.61       | 0.06 $\checkmark$ | 0.22              | 0.07 $\checkmark$ |
| Qwen3-8B     | 0.86      | 0.71 | 0.66  | 0.37       | 0.04 $\checkmark$ | 0.15              | 0.01 $\checkmark$ |
| Qwen3-0.6B   | 0.92      | 0.63 | 0.50  | 0.56       | 0.01 $\checkmark$ | 0.08 $\checkmark$ | 0.10 $\checkmark$ |
| Mistral-7B   | 0.89      | 0.82 | 0.82  | 0.23       | 0.00 $\checkmark$ | 0.16              | 0.04 $\checkmark$ |
| Equiv. count | 0/8       | 0/8  | 0/8   | 0/8        | 6/8               | 4/8               | 8/8               |

### u.1 TOST Bootstrap Stability

To verify that the equivalence conclusions in §u are not sensitive to sample composition, we assess their stability under bootstrap resampling. For each of four representative features, we run 1,000 bootstrap iterations (base seed = 42): each iteration independently resamples with replacement from the gen-1 and gen-2+ pools, recomputes Cohen’s  $d$  and the TOST  $p$ -value (equivalence margin  $|d| < 0.20$ ).

All three near-stationary features achieve a 100% rejection rate across 1,000 bootstrap iterations, and their effect-size 95% CIs fall entirely within the  $\pm 0.20$  equivalence margin. Entropy std, included as an intermediate control, is never rejected (0%), with its CI  $[0.435, 0.479]$  lying entirely outside the margin. No feature exhibits boundary-crossing: the classification of each feature is unanimous across

Table 34: Bootstrap stability of TOST equivalence ( $B=1,000$ , margin  $|d|<0.20$ ). Rejection rate = fraction of bootstrap iterations where TOST rejects non-equivalence at  $\alpha=0.05$ .

| Feature     | Rejection rate   | Median $d$ | $d$ 95% CI      |
|-------------|------------------|------------|-----------------|
| entropy_max | 100% (1000/1000) | 0.062      | [0.037, 0.088]  |
| entropy_min | 100% (1000/1000) | -0.016     | [-0.042, 0.011] |
| tail_decay  | 100% (1000/1000) | 0.088      | [0.065, 0.111]  |
| entropy_std | 0% (0/1000)      | 0.458      | [0.435, 0.479]  |

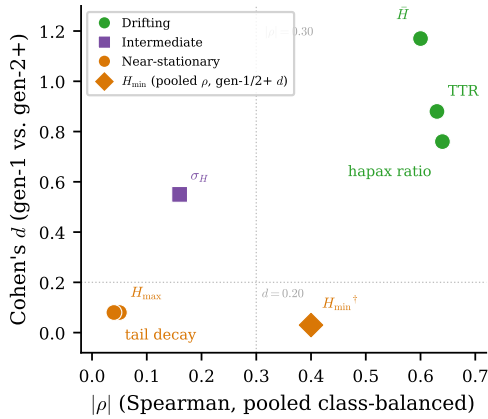


Figure 17: DDR partition in pooled ( $|\rho|$ , Cohen’s  $d$ ) space. Drifting features (red) cluster at high  $|\rho|$  and  $d$ ; near-stationary features (blue) cluster near the origin; entropy std (purple) occupies a sole intermediate position. Dashed lines: TOST equivalence bound ( $d=0.20$ ) and  $|\rho|=0.30$ . Diamond marker: entropy min (pooled  $|\rho|=0.40$  reflects the gen-0/gen-1 boundary; gen-1 vs. gen-2+  $d=0.03$ ).

all bootstrap samples. The equivalence conclusions are robust to resampling variability.

## v DDR Partition Validation

### Data-Driven Clustering

To assess the DDR partition structure and determine whether the drifting/non-drifting split is data-driven, we apply  $k$ -means clustering ( $k \in \{2, 3, 4, 5\}$ ) to the seven features in standardized ( $|\rho|, d$ ) space (Table 35).

Table 35:  $k$ -means clustering of 7 features in ( $|\rho|, d$ ) space.

| $k$ | Silhouette   | Clusters                                                                                                   |
|-----|--------------|------------------------------------------------------------------------------------------------------------|
| 2   | <b>0.660</b> | $\{\bar{H}, \text{TTR}, \text{hapax}\} \mid \{\sigma_H, H_{\min}, H_{\max}, \text{tail}\}$                 |
| 3   | 0.532        | $\{\bar{H}, \text{TTR}, \text{hapax}\} \mid \{\sigma_H, H_{\max}, \text{tail}\} \mid \{H_{\min}\}$         |
| 4   | 0.527        | $\{\bar{H}, \text{TTR}, \text{hapax}\} \mid \{H_{\max}, \text{tail}\} \mid \{\sigma_H\} \mid \{H_{\min}\}$ |

The primary drifting/non-drifting split ( $k=2$ , silhouette = 0.660) is the strongest cluster structure.

At  $k=3$ , entropy min forms its own cluster rather than joining the near-stationary group because its pooled  $|\rho|=0.40$  (driven by the gen-0/gen-1 boundary) places it far from other near-stationary features in  $\rho$  space. The 3/1/3 DDR partition resolves this by using the gen-1 vs. gen-2+ Cohen’s  $d$  ( $d=0.03$ ) to place entropy min in the near-stationary group, a domain-informed refinement that unsupervised clustering cannot recover from pooled statistics alone. At  $k=4$  (silhouette = 0.527), entropy std separates from the near-stationary group, consistent with its intermediate DDR classification.

**Exhaustive partition enumeration.** To confirm that the three-vs-four split is optimal rather than incidental, we enumerate all  $\binom{7}{3}=35$  possible ways to assign three features to a “drifting” group and the remaining four to “non-drifting,” computing the silhouette score for each partition in standardized ( $|\rho|, d$ ) space. The observed DDR partition ( $\{\text{entropy mean, TTR, hapax ratio}\}$ ) achieves the highest silhouette (0.660), with a gap of 0.18 to the second-ranked split; 20 of 35 partitions produce negative silhouette scores.

**Per-generator replication.** To rule out the possibility that the pooled silhouette is an artifact of mixing heterogeneous generators, we repeat the ( $|\rho|, d$ ) clustering independently for each of the eight generators. Every generator recovers the identical  $\{\bar{H}, \text{TTR, hapax}\}$  vs.  $\{\sigma_H, H_{\min}, H_{\max}, \text{tail}\}$  partition, with silhouette scores ranging from 0.505 (Pythia-1B) to 0.760 (Mistral-7B-v0.3; mean = 0.651). No generator falls below the 0.50 moderate-structure threshold. The two-cluster structure is a per-generator property rather than a pooling artifact.

### TOST Equivalence Bound Sensitivity

Table 36 shows the fraction of generators for which each feature passes the TOST equivalence test (gen-1 vs. gen-2+,  $\alpha=0.05$ ) at three equivalence bounds  $\varepsilon$ .

Table 36: TOST equivalence pass rate across 8 generators at varying  $\varepsilon$ .

| DDR Zone     | Feature               | $\varepsilon=0.15$ | $\varepsilon=0.20$ | $\varepsilon=0.25$ |
|--------------|-----------------------|--------------------|--------------------|--------------------|
| Drifting     | $\bar{H}$ (ent. mean) | 0/8                | 0/8                | 0/8                |
|              | TTR                   | 0/8                | 0/8                | 0/8                |
|              | hapax ratio           | 0/8                | 0/8                | 0/8                |
| Intermediate | $\sigma_H$ (ent. std) | 0/8                | 0/8                | 0/8                |
| Near-stat.   | $H_{\min}$ (ent. min) | 7/8                | 8/8                | 8/8                |
|              | $H_{\max}$ (ent. max) | 6/8                | 6/8                | 6/8                |
|              | tail decay            | 4/8                | 4/8                | 7/8                |

The partition is robust to the equivalence bound: drifting features reject equivalence at all  $\varepsilon$  for all generators, while entropy min passes even at the stricter bound  $\varepsilon=0.15$  (7/8 generators). Tail decay is the most  $\varepsilon$ -sensitive near-stationary feature, passing 4/8 at  $\varepsilon=0.20$  but 7/8 at  $\varepsilon=0.25$ ; the four non-passing generators show  $|d|=0.15\text{--}0.22$ , still well below the drifting threshold ( $d > 0.50$ ). Entropy max is stable across all three bounds (6/8), with the two failures (Pythia-1B  $d=0.21$ , Pythia-2.8B  $d=0.25$ ) reflecting same-family reference affinity with the Pythia-410M reference model.

**TOST margin rationale.** We adopt  $\varepsilon=0.20$  as the primary equivalence bound, corresponding to Cohen’s “small effect” threshold ( $d=0.20$ ; Cohen, 1988). This choice reflects a principled smallest effect size of interest (SESOI): effects below  $d=0.20$  are conventionally negligible and unlikely to carry practical discriminative value for depth classification. The bound is conservative relative to the drifting–stationary gap ( $d_{\min}^{\text{drift}}=0.74$  vs.  $\varepsilon=0.20$ , a  $3.7\times$  margin); the equivalence test thus has adequate power to confirm near-stationarity without risk of misclassifying genuinely drifting features. As Table 36 shows, conclusions are qualitatively identical at  $\varepsilon=0.15$  and  $\varepsilon=0.25$ ; the partition is insensitive to the specific bound choice. LOFO analysis independently corroborates this boundary: near-stationary features contribute  $<2$  pp marginal importance (Appendix f), empirically validating  $\varepsilon=0.20$  as a task-relevant threshold.

**Threshold sensitivity.** To address concerns about the sensitivity of the drifting-boundary threshold, we sweep  $|d|$  from 0.10 to 0.80 in increments of 0.05 and record the resulting binary partition (drifting vs. stationary) along with the silhouette score in standardized  $(|\rho|, d)$  space. Across the entire range, only *three* distinct partitions emerge, reflecting natural gaps between feature clusters rather than threshold artifacts. The three-feature drifting set (entropy mean, TTR, hapax ratio) is stable across  $|d| \in [0.50, 0.75]$  with a silhouette of 0.660 (the peak value in the sweep). Including the intermediate feature entropy std ( $d=0.46$ ) yields a 4/3 partition stable across  $|d| \in [0.10, 0.45]$  (silhouette = 0.451). The partition used in our analysis corresponds to the peak-silhouette region, confirming that the drifting/stationary boundary reflects a natural gap in the  $(|\rho|, d)$  landscape rather than an artifact of a particular threshold choice.

**Partition stability across  $\varepsilon$  thresholds.** The DDR partition classifies features as non-stationary ( $d \geq \varepsilon$ ) or stationary ( $d < \varepsilon$ ) based on pooled Cohen’s  $d$  (§4.2). Table 37 reports the critical  $\varepsilon$  at which each feature’s classification changes; the three drifting and three near-stationary features maintain their classification for all  $\varepsilon \in [0.10, 0.75]$ .

Table 37: Partition stability under varying  $\varepsilon$ . Critical threshold: the  $\varepsilon$  above which a feature is reclassified as stationary. The drifting/near-stationary split is invariant for  $\varepsilon \in [0.10, 0.75]$ ; only entropy std ( $d=0.46$ ) changes at  $\varepsilon > 0.46$ .

| DDR Zone     | Feature               | Pooled $d$ | Stationary at $\varepsilon >$ |
|--------------|-----------------------|------------|-------------------------------|
| Drifting     | $\bar{H}$ (ent. mean) | 1.05       | 1.05                          |
|              | TTR                   | 0.81       | 0.81                          |
|              | hapax ratio           | 0.74       | 0.74                          |
| Intermediate | $\sigma_H$ (ent. std) | 0.46       | 0.46                          |
| Near-stat.   | $H_{\max}$ (ent. max) | 0.06       | 0.06                          |
|              | tail decay            | 0.09       | 0.09                          |
|              | $H_{\min}$ (ent. min) | 0.016      | 0.016                         |

The 3/1/3 partition is stable across  $\varepsilon \in [0.10, 0.75]$ : the three drifting features remain non-stationary at all  $\varepsilon < 0.74$ , the three near-stationary features remain stationary at all  $\varepsilon > 0.09$ , and only entropy std changes classification (becoming stationary at  $\varepsilon > 0.46$ ). It is the sole threshold-sensitive feature. The minimum gap between adjacent zone boundaries ( $d_{\text{hapax}} - d_{\sigma_H} = 0.28$ ) exceeds typical effect-size estimation uncertainty and provides a margin against sampling noise.

### Holdout Generator Split Validation

To confirm the DDR partition is not an artifact of the specific generator pool, we split the eight generators into disjoint selection (4) and validation (4) sets, compute per-subset pooled  $(|\rho|, d)$ , and check whether all seven features receive the same zone assignment in both subsets. We evaluate two complementary splits to avoid dependence on a single partition: Split A selects {Pythia-2.8B, Qwen3-1.7B, OLMo, Mistral} (validation: Pythia-1B, Qwen3-0.6B, Qwen3-8B, Llama); Split B selects {Pythia-1B, Qwen3-0.6B, OLMo, Mistral} (validation: Pythia-2.8B, Qwen3-1.7B, Qwen3-8B, Llama).

Under simple joint  $(|\rho|, d)$  thresholds, both splits produce a 3/2/2 partition: entropy min is classified as a second intermediate feature due to its elevated pooled  $|\rho|$  (0.36–0.39). However, entropy min’s binary (gen-1 vs. gen-2+)  $d$  values are uniformly below 0.06 across all four split  $\times$  subset combina-

Table 38: Holdout split validation of the DDR partition. Selection (sel) and validation (val)  $|\rho|$  and  $d$  for two disjoint 4/4 generator splits. Zone assignment is identical across all four subsets for every feature.

| Zone         | Feature               | Split A      |              |                | Split B      |              |                |
|--------------|-----------------------|--------------|--------------|----------------|--------------|--------------|----------------|
|              |                       | sel $ \rho $ | val $ \rho $ | $\Delta \rho $ | sel $ \rho $ | val $ \rho $ | $\Delta \rho $ |
| Drifting     | $\bar{H}$ (ent. mean) | 0.60         | 0.58         | 0.02           | 0.60         | 0.58         | 0.02           |
|              | TTR                   | 0.61         | 0.59         | 0.02           | 0.60         | 0.59         | 0.01           |
|              | hapax ratio           | 0.62         | 0.60         | 0.02           | 0.62         | 0.61         | 0.01           |
| Intermediate | $\sigma_H$ (ent. std) | 0.19         | 0.21         | 0.02           | 0.22         | 0.17         | 0.05           |
| Near-stat.   | $H_{\min}$ (ent. min) | 0.39         | 0.36         | 0.03           | 0.38         | 0.37         | 0.01           |
|              | $H_{\max}$ (ent. max) | 0.05         | 0.05         | 0.00           | 0.04         | 0.06         | 0.02           |
|              | tail decay            | 0.10         | 0.08         | 0.02           | 0.11         | 0.07         | 0.04           |

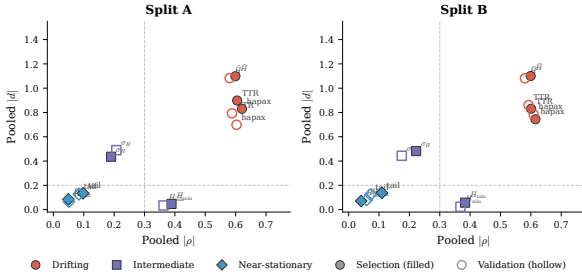


Figure 18: Holdout split validation in  $(|\rho|, d)$  space. Filled markers: selection set; hollow: validation set. Zone assignments are identical across subsets; selection/validation pairs cluster tightly for all features.

tions, consistent with near-stationarity on the task-relevant metric; its pooled  $|\rho|$  reflects change concentrated at the gen-0/gen-1 boundary rather than sustained drift. Five of seven features are unambiguously classified under any reasonable criterion; entropy min’s boundary status is resolved by the binary-task  $d$  criterion.

Selection and validation subsets produce identical partitions in both splits (Table 38; Figure 18). The maximum  $\Delta|\rho|$  between selection and validation is 0.05 (entropy std, Split B), and the minimum gap between adjacent zones remains  $>0.15$  in all subsets.

### Heteroscedasticity Sensitivity

Levene tests reject equal variance between gen-1 and gen-2+ for drifting features ( $p < 10^{-10}$ ; variance ratios 1.86–2.21). To verify that the DDR partition is robust to this heteroscedasticity, we supplement pooled Cohen’s  $d$  with Glass’s  $\Delta$  (using the gen-1 standard deviation as denominator) and Cliff’s  $\delta$  (a non-parametric ordinal effect size; Table 39).

Effect sizes in Table 39 are computed on the binary task (gen-1 vs. gen-2+); the main text (§4.2) reports 3-class pooled  $d$  values (e.g., hapax ratio  $d=0.74$ ), which are slightly larger because the

Table 39: Effect size comparison (gen-1 vs. gen-2+,  $N = 29,095$ ). Glass’s  $\Delta$  uses the gen-1 SD as denominator; Cliff’s  $\delta$  thresholds follow Romano et al. (2006):  $|\delta| < 0.147$  negligible,  $< 0.33$  small,  $< 0.474$  medium, otherwise large.

| Zone         | Feature     | $ d $ | $ \Delta $ | $ \delta $         |
|--------------|-------------|-------|------------|--------------------|
| Drifting     | ent. mean   | 1.054 | 1.410      | 0.587              |
|              | TTR         | 0.814 | 1.116      | 0.510              |
|              | hapax ratio | 0.735 | 0.936      | 0.528              |
| Intermediate | ent. std    | 0.457 | 0.605      | 0.295              |
| Near-stat.   | tail decay  | 0.087 | 0.099      | 0.072              |
|              | ent. max    | 0.062 | 0.065      | 0.021              |
|              | ent. min    | 0.016 | 0.015      | 0.354 <sup>†</sup> |

gen-0/gen-1 contrast contributes additional separation. Glass’s  $\Delta$  amplifies drifting effects (1.41 vs. 1.05 for entropy mean) because gen-2+ variance exceeds gen-1 variance (ratio 2.08); pooled Cohen’s  $d$  is therefore conservative. Near-stationary features remain negligible under both parametric metrics ( $|\Delta| \leq 0.099$ ). The minimum drifting-to-maximum-near-stationary gap is  $9.5\times$  under Glass’s  $\Delta$  ( $0.936/0.099$ ); heteroscedasticity does not erode the partition. Cliff’s  $\delta$  classifies all drifting features as “large” ( $|\delta| > 0.474$ ) and entropy std as “small” ( $|\delta|=0.295$ ), consistent with the parametric partition.

<sup>†</sup>Entropy min’s elevated Cliff’s  $\delta$  (0.354, “medium”) reflects distributional shape asymmetry between gen-1 and gen-2+ that rank-based statistics detect but location measures do not (Cohen’s  $d=0.016$ , Glass’s  $\Delta=0.015$ ). This is consistent with its documented boundary status (§4.2): the gen-0/gen-1 transition produces rank-order shifts that persist into deeper generations without affecting the mean.

### w Generator Subset Summary

Table 40 summarizes the generator subsets used across experiments.

Table 40: Generator subsets by experiment type.

| Experiment                      | $N$ | Generators                                             |
|---------------------------------|-----|--------------------------------------------------------|
| Main LOGO (Table 2)             | 8   | All: Py-2.8B/1B, OLMo, Qw-1.7B/0.6B/8B, Llama, Mistral |
| Cross-model chains (Table 31)   | 5   | Py-2.8B/1B, OLMo, Qw-1.7B/0.6B                         |
| QLoRA (Table 7)                 | 6   | Py-2.8B/1B, OLMo, Qw-1.7B/0.6B/8B                      |
| Temperature ablation (Table 14) | 4   | Py-2.8B, OLMo, Qw-1.7B/0.6B                            |
| Prompt-type (Appendix c)        | 1   | Py-2.8B only                                           |
| Cross-domain (Table 18)         | 3   | Py-1B, Qw-1.7B, Mistral                                |
| Instruct (Table 19)             | 3   | OLMo, Qw-1.7B, Llama                                   |
| Downstream (Appendix m)         | 5   | Py-2.8B/1B, OLMo, Qw-1.7B/0.6B                         |

The 5-generator set was the initial development subset; Qwen3-8B was added for scale evaluation, and Llama-3.2-3B and Mistral-7B-v0.3 were added

Table 41: Example text chain from seed 120 (Pythia-2.8B, continuation prompt). Gen-0 is the original Wikipedia passage; gen-1 and gen-2 are successive LLM continuations. Text is truncated for space.

| Gen-0 (Wikipedia seed)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Aquarius is an equatorial constellation of the zodiac, between Capricornus and Pisces. Its name is Latin for “water-carrier” or “cup-carrier”. Aquarius is one of the oldest of the recognized constellations along the zodiac (the Sun’s apparent path). It was one of the 48 constellations listed by the 2nd century astronomer Ptolemy, and it remains one of the 88 modern constellations. It is found in a region often called the Sea due to its profusion of constellations with watery associations such as Cetus the whale, Pisces the fish, and Eridanus the river. [...] |
| Gen-1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Babylonian astrology and was also depicted in ancient Egyptian art. In Greek mythology, Aquarius is the son of Poseidon and Gaia. He is also the god of the sea and the sky, the god of travelers, and the guardian of the underworld. In Roman mythology, Aquarius is the son of Saturn and the nymph Egeria, and the god of the sea. In Jewish mythology, Aquarius is the son of the God of the Air and the Goddess of the Moon. In Chinese mythology, Aquarius is the son of the God of the Sky and the Goddess of the Earth. [...]                                               |
| Gen-2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| God of the Sky and the Goddess of the Earth. In the ancient Greek mythology, Aquarius is the son of the God of the Sky and the Goddess of the Earth. In the Hindu mythology, Aquarius is the son of the God of the Sky and the Goddess of the Earth. In the ancient Egyptian mythology, Aquarius is the son of the God of the Sky and the Goddess of the Earth. [...]                                                                                                                                                                                                                |

last for family-diversity testing. Subset differences reflect chronological expansion rather than selective reporting.

## x Qualitative Examples

Table 41 presents an example generation chain from seed 120 (Pythia-2.8B, continuation prompt), illustrating how text evolves across recursive generation steps. Table 42 reports the corresponding 7-feature values. Central-tendency features ( $\bar{H}$ , TTR, hapax ratio) drift monotonically while shape features ( $H_{\max}$ , tail decay) remain near-stationary, consistent with the differential depth response pattern.

## y Cross-Generator Transfer Analysis

To isolate the depth signal from generator identity, we train a binary classifier (gen-1 vs. gen-2+) on each of the eight generators independently and evaluate on all others. Table 43 reports the full  $8 \times 8$

Table 42: Feature values for the chain in Table 41. Drifting features ( $\bar{H}$ , TTR, hapax) show monotonic degradation; shape features ( $H_{\max}$ , tail decay) remain stable (range  $<2\%$  and  $<4\%$  respectively).

| Feature                  | Gen-0    | Gen-1    | Gen-2    |
|--------------------------|----------|----------|----------|
| $\bar{H}$ (entropy mean) | 2.720    | 2.206    | 0.780    |
| $\sigma_H$ (entropy std) | 1.772    | 1.816    | 1.517    |
| $H_{\min}$ (entropy min) | 0.002    | 0.005    | 0.000    |
| $H_{\max}$ (entropy max) | 7.000    | 7.016    | 6.926    |
| tail decay               | −0.00227 | −0.00231 | −0.00223 |
| TTR                      | 0.659    | 0.272    | 0.081    |
| hapax ratio              | 0.561    | 0.183    | 0.005    |

transfer matrix; Figure 3 (main text) visualizes it as a heatmap. Note that these classifiers are trained with a binary objective, whereas the LOGO classifiers in Table 2 are trained on a 3-class task and evaluated on the binary subtask; the two setups are therefore not directly comparable.

Table 43: Cross-generator binary F1 transfer matrix (gen-1 vs. gen-2+). Each row trains on one generator; each column evaluates on another. Diagonal entries (underlined) are within-generator. All 56 off-diagonal pairs exceed chance (0.5). Classifiers use the same 7-feature XGBoost configuration as LOGO but are trained with a binary objective.

| Train \ Test | Py-2.8B      | Py-1B        | Qw-0.6B      | Qw-1.7B      | Qw-8B        | OLMo-1B      | LI-3B        | Mi-7B        |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Py-2.8B      | <u>0.985</u> | 0.873        | 0.823        | 0.833        | 0.741        | 0.829        | 0.815        | 0.712        |
| Py-1B        | 0.834        | <u>0.987</u> | 0.828        | 0.821        | 0.704        | 0.778        | 0.764        | 0.620        |
| Qw-0.6B      | 0.786        | 0.846        | <u>0.991</u> | 0.851        | 0.786        | 0.795        | 0.760        | 0.680        |
| Qw-1.7B      | 0.794        | 0.839        | 0.859        | <u>0.989</u> | 0.790        | 0.817        | 0.798        | 0.717        |
| Qw-8B        | 0.811        | 0.836        | 0.853        | 0.855        | <u>0.984</u> | 0.848        | 0.839        | 0.790        |
| OLMo-1B      | 0.811        | 0.827        | 0.837        | 0.852        | 0.803        | <u>0.988</u> | 0.828        | 0.763        |
| LI-3B        | 0.856        | 0.860        | 0.845        | 0.851        | 0.798        | 0.851        | <u>0.987</u> | 0.787        |
| Mi-7B        | 0.834        | 0.837        | 0.837        | 0.845        | 0.817        | 0.849        | 0.851        | <u>0.969</u> |

All 56 off-diagonal transfer pairs yield F1 above chance, with a mean of 0.810 (95% bootstrap CI [0.796, 0.822]; one-sample  $t = 46.58$ ,  $p < 10^{-10}$ ). Transfer is strongest within families (e.g., Py-2.8B→Py-1B: 0.87) and weakest for Py-1B→Mi-7B (0.62), consistent with the scale and reference-affinity effects discussed in §4.2. These results confirm that depth signatures generalize across architectures, parameter scales, and model families. Generator identity is not a primary confound.

## z Supplementary DDR Characterization

### z.1 Partition Stability Across Temperatures

To assess whether the DDR three-way partition generalizes beyond the default temperature ( $\tau=0.7$ ), we computed per-feature Cohen’s  $d$  (generation 1 vs. generation 2+) at  $\tau \in \{0.5, 0.7, 0.9\}$  for all four generators in the temperature ablation (12 conditions total; Table 44).

The partition anchors are robust: entropy mean

Table 44: Mean  $|d|$  (gen-1 vs. gen-2+) across 4 generators at each temperature. Partition labels: Drifting ( $|d| > 0.5$ ), Intermediate ( $0.1 < |d| < 0.5$ ), near-Stationary ( $|d| < 0.1$ ). Boldface marks features that shift from their  $\tau=0.7$  partition.

| Feature      | $\tau=0.5$              | $\tau=0.7$              | $\tau=0.9$              | Group        |
|--------------|-------------------------|-------------------------|-------------------------|--------------|
| Entropy mean | 1.39 <sup>D</sup>       | 1.10 <sup>D</sup>       | 0.58 <sup>D</sup>       | Drifting     |
| TTR          | 1.00 <sup>D</sup>       | 0.85 <sup>D</sup>       | <b>0.45<sup>I</sup></b> | Drifting     |
| Hapax ratio  | 0.87 <sup>D</sup>       | 0.75 <sup>D</sup>       | <b>0.42<sup>I</sup></b> | Drifting     |
| Entropy std  | <b>0.84<sup>D</sup></b> | 0.52 <sup>D</sup>       | <b>0.20<sup>I</sup></b> | Intermediate |
| Entropy max  | 0.06 <sup>S</sup>       | 0.09 <sup>S</sup>       | <b>0.11<sup>I</sup></b> | Stationary   |
| Tail decay   | 0.07 <sup>S</sup>       | <b>0.12<sup>I</sup></b> | <b>0.15<sup>I</sup></b> | Stationary   |
| Entropy min  | 0.04 <sup>S</sup>       | 0.06 <sup>S</sup>       | 0.04 <sup>S</sup>       | Stationary   |

remains drifting in all 12 conditions, TTR in 10/12, and hapax ratio in 9/12; entropy min and entropy max stay near-stationary in 10/12 and 9/12 conditions, respectively. The two boundary features are more temperature-sensitive: entropy std is drifting at  $\tau=0.5$  for all generators but falls to intermediate at  $\tau=0.9$  for three of four, while tail decay fluctuates between near-stationary and intermediate across generators. Effect sizes compress monotonically with temperature: for the fixed drifting set (entropy mean, TTR, hapax ratio), mean  $|d|$  drops from 1.09 at  $\tau=0.5$  to 0.90 at  $\tau=0.7$  to 0.48 at  $\tau=0.9$ , shifting individual features across partition boundaries at high temperature without altering the qualitative hierarchy: lexical diversity features drift fastest, entropy extrema remain stable across all conditions tested.

## z.2 Per-Family DDR Consistency

Table 45: Intra-family consistency of DDR patterns. Each cell shows the mean Spearman  $|\rho|$  (depth, feature) across generators within a family ( $\pm$  SD for multi-generator families). Drifting features exhibit consistently high  $|\rho|$  while near-stationary features remain low across all families.  $H_{\min}$  is a *boundary feature*: its elevated  $|\rho| \approx 0.40$  reflects a large gen-0  $\rightarrow$  gen-1 transition, whereas its Cohen’s  $d \approx 0$  for gen-1 / gen-2+ confirms near-stationarity within the binary classification scope (§4.2).

| Family     | $n$ | DRIFTING          |                 |                 | INT.            | NEAR-STATIONARY  |                  |                       |
|------------|-----|-------------------|-----------------|-----------------|-----------------|------------------|------------------|-----------------------|
|            |     | $H_{\text{mean}}$ | TTR             | Hapax           |                 | $H_{\text{std}}$ | $H_{\text{max}}$ | $\beta_{\text{tail}}$ |
| Pythia     | 2   | 0.69 $\pm$ 0.03   | 0.65 $\pm$ 0.02 | 0.66 $\pm$ 0.01 | 0.20 $\pm$ 0.06 | 0.02 $\pm$ 0.02  | 0.49 $\pm$ 0.02  | 0.08 $\pm$ 0.02       |
| OLMo       | 1   | 0.64              | 0.70            | 0.71            | 0.30            | 0.07             | 0.39             | 0.14                  |
| Qwen       | 3   | 0.55 $\pm$ 0.07   | 0.60 $\pm$ 0.06 | 0.62 $\pm$ 0.06 | 0.16 $\pm$ 0.08 | 0.09 $\pm$ 0.00  | 0.36 $\pm$ 0.05  | 0.08 $\pm$ 0.01       |
| Llama      | 1   | 0.63              | 0.69            | 0.70            | 0.16            | 0.09             | 0.40             | 0.06                  |
| Mistral    | 1   | 0.54              | 0.60            | 0.61            | 0.06            | 0.09             | 0.40             | 0.03                  |
| All 8 gen. | 8   | 0.60 $\pm$ 0.08   | 0.63 $\pm$ 0.05 | 0.65 $\pm$ 0.05 | 0.18 $\pm$ 0.08 | 0.07 $\pm$ 0.03  | 0.40 $\pm$ 0.06  | 0.08 $\pm$ 0.03       |

## z.3 Per-Feature Convergence Dynamics

Figure 6 (main text) visualizes per-feature convergence trajectories across generation depths 0–4, averaged over 8 generators; each feature is expressed as Cohen’s  $d$  relative to depth 0. Drifting features (hapax ratio, TTR, entropy mean) and the intermediate feature (entropy std) diverge monotonically; lexical diversity features (TTR, hapax ratio) form a fast-convergence group ( $|d| > 2$  by depth 2), while entropy mean and entropy std drift more slowly. Near-stationary features (entropy max, tail decay, entropy min) remain within  $|d| < 0.25$  across all depths, consistent with location stability. Shaded bands indicate  $\pm 1$  standard deviation across generators.

## z.4 DDR Causality Analysis

This appendix validates the causal predictive power of DDR features through three complementary analyses: permutation importance, Granger causality, and feature intervention.

**Permutation Importance.** Table 46 reports the mean F1 drop when each feature’s values are randomly shuffled across the test set (baseline F1 = 0.771). Drifting features dominate: their average F1 drop (0.187) is  $5.5\times$  larger than the near-stationary average (0.034). Location-shift features, not shape features, drive depth classification.

Table 46: Permutation importance (8-gen LOGO). Each feature is shuffled independently; the resulting F1 drop quantifies marginal importance. Drifting features average 0.187 vs. near-stationary 0.034 ( $5.5\times$ ).

| Feature      | Group        | Mean F1 Drop | Rank |
|--------------|--------------|--------------|------|
| hapax_ratio  | Drifting     | 0.235        | 1    |
| entropy_mean | Drifting     | 0.170        | 2    |
| ttr          | Drifting     | 0.156        | 3    |
| entropy_max  | Near-stat.   | 0.070        | 4    |
| entropy_std  | Intermediate | 0.057        | 5    |
| entropy_min  | Near-stat.   | 0.019        | 6    |
| tail_decay   | Near-stat.   | 0.014        | 7    |

**Granger Causality.** Table 47 reports Fisher-combined  $p$ -values from panel Granger causality tests (lag = 1) across 8 generators ( $N = 29,095$ ). Six of seven features exhibit significant Granger causality ( $p < 0.001$ ), meaning past feature values predict future depth beyond what past depth alone explains. The sole exception is entropy\_min ( $p = 0.375$ ). This provides additional evidence for its near-stationary classification: it carries neither

large marginal importance (Table 46, rank 7) nor temporal predictive direction.

Table 47: Granger causality (Fisher combined across 8 generators, lag = 1,  $N = 29,095$ ).  $\Delta R^2$  is reported where available.

| Feature      | Combined $p$ | $\Delta R^2$ | Significant |
|--------------|--------------|--------------|-------------|
| entropy_std  | < 0.001      | 0.031        | Yes         |
| hapax_ratio  | < 0.001      | —            | Yes         |
| ttr          | < 0.001      | —            | Yes         |
| entropy_mean | < 0.001      | —            | Yes         |
| entropy_max  | < 0.001      | —            | Yes         |
| tail_decay   | < 0.001      | —            | Yes         |
| entropy_min  | 0.375        | —            | No          |

### Key Findings.

- Drifting dominance.** The  $5.5\times$  permutation-importance gap confirms that drifting features (entropy\_mean, ttr, hapax\_ratio) are the primary depth signal carriers, consistent with the DDR partition derived from convergence analysis (§4.2).
- Entropy\_min lacks causal direction.** Beyond its small effect size (F1 drop = 0.019), entropy\_min is the only feature without significant Granger causality ( $p = 0.375$ ). This strengthens the near-stationary classification: the feature is neither marginally important nor temporally predictive.
- Entropy\_max asymmetry.** Entropy\_max ranks fourth in permutation importance (F1 drop = 0.070) but produces the largest degradation in the counterfactual intervention ( $-2.0$  pp). It provides unique signal not redundant with other features; its causal contribution exceeds its marginal importance.
- Mistral heterogeneity.** In per-generator Granger tests, Mistral-7B shows non-significant hapax\_ratio and ttr ( $p > 0.4$ ), though Fisher combination across generators remains significant. This generator-specific attenuation partly explains Mistral’s lowest LOGO F1 (0.605) and suggests that DDR depth signatures are weaker in models with heavier repetition penalties.

### z.5 Variance Decomposition

To quantify the mechanistic structure underlying DDR, we decompose each feature’s variance into depth, generator, interaction, and residual components via two-way ANOVA (Type II; Table 48).

The depth factor explains 34–37% of location-feature variance but <1% of shape-feature vari-

Table 48: Two-way ANOVA variance decomposition (classical  $\eta^2$ ; proportions of total SS) for depth  $\times$  generator ( $N = 37,058$ ). Location features show dominant depth effects ( $\eta_{\text{depth}}^2 = 0.34\text{--}0.37$ ); shape features are depth-invariant ( $\eta_{\text{depth}}^2 < 0.01$ ). All effects  $p < 10^{-6}$ .

| Feature                        | $\eta_{\text{depth}}^2$ | $\eta_{\text{gen}}^2$ | $\eta_{\text{int}}^2$ | $\eta_{\text{resid}}^2$ |
|--------------------------------|-------------------------|-----------------------|-----------------------|-------------------------|
| <i>Location (drifting)</i>     |                         |                       |                       |                         |
| Ent. Mean                      | <b>0.341</b>            | 0.022                 | 0.010                 | 0.627                   |
| TTR                            | <b>0.370</b>            | 0.024                 | 0.007                 | 0.599                   |
| Hapax Ratio                    | <b>0.364</b>            | 0.017                 | 0.006                 | 0.612                   |
| <i>Shape (near-stationary)</i> |                         |                       |                       |                         |
| Ent. Max                       | 0.007                   | 0.002                 | 0.003                 | 0.988                   |
| Ent. Min                       | 0.002                   | 0.001                 | 0.001                 | 0.996                   |
| Tail Decay                     | 0.002                   | 0.023                 | 0.009                 | 0.967                   |
| <i>Intermediate</i>            |                         |                       |                       |                         |
| Ent. Std                       | 0.054                   | 0.020                 | 0.011                 | 0.915                   |

ance, a  $>48\times$  ratio that quantifies the DDR partition as a systematic depth  $\times$  feature-type interaction rather than a post-hoc grouping. Generator identity contributes comparably across both groups ( $\eta_{\text{gen}}^2 = 0.02$ ); the asymmetry is specific to depth, not generator. Interaction terms are uniformly small ( $\eta_{\text{int}}^2 < 0.011$ ); the depth effect on location features operates consistently across generators.

Exponential convergence fitting ( $f(d) = a \cdot e^{-\lambda d} + c$ ) on per-generator depth trajectories corroborates this: location features converge with moderate rates ( $\lambda = 0.40 \pm 0.26$ ,  $R^2 = 0.99$ ), consistent with contraction toward a generator-specific fixed point under iterated autoregressive generation, while shape features show degenerate fits ( $\lambda = 5.3 \pm 8.4$ ,  $R^2 = 0.73$ ) reflecting their flat trajectories.

### z.6 Decoding Strategy Robustness

To verify that DDR shape stability is not an artifact of nucleus sampling ( $p = 0.9$ ,  $\tau = 0.7$ ), we repeat the generation chain experiment under two alternative decoding strategies: greedy decoding ( $\tau = 0$ ) and top- $k$  sampling ( $k = 50$ ,  $\tau = 0.7$ ). All eight generators spanning five architectural families produce  $\sim 1,000$  chains of depth 0–4 per strategy, yielding  $\sim 5,000$  texts per generator–strategy combination.

Table 49 reports Cohen’s  $d$  (gen-1 vs. gen-2+) for the three shape features across strategies; Table 50 provides the corresponding values for drifting and intermediate features.

Under greedy decoding ( $\tau = 0$ ), shape features remain near-stationary across all eight generators (mean  $|d| = \mathbf{0.05}$ , max **0.14**), while core drifting features maintain large effects (mean  $d = \mathbf{0.94}$ ). This confirms that DDR is an empirical regular-

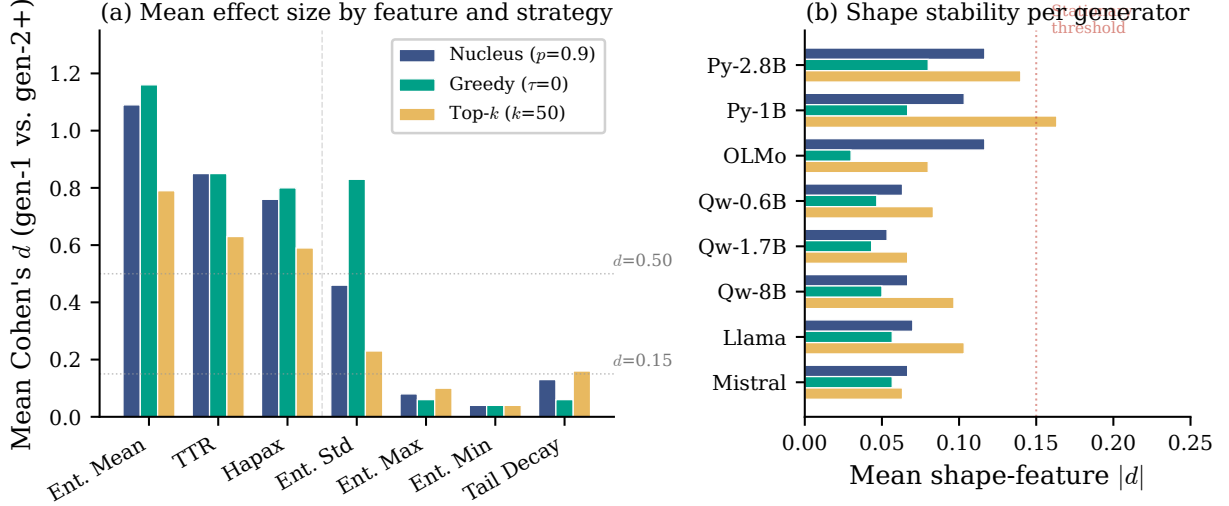


Figure 19: Decoding strategy robustness. (a) Mean Cohen’s  $d$  (gen-1 vs. gen-2+) across three decoding strategies. Drifting features maintain large  $d$  across all strategies; shape features remain below the stationarity threshold ( $d < 0.15$ ). Reference lines mark drifting ( $d=0.50$ ) and stationary ( $d=0.15$ ) thresholds. (b) Per-generator mean shape-feature  $|d|$ : all generators stay near-stationary under all three strategies.

Table 49: Shape-feature Cohen’s  $d$  (gen-1 vs. gen-2+) under three decoding strategies per generator. Values below 0.20 indicate negligible drift. Shape features are expected to remain near-stationary across strategies if DDR is decoding-agnostic.

| Family  | Generator | entropy_max         |                     |                     | entropy_min         |                     |                     | tail_decay          |                     |                     |
|---------|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|         |           | Nuc.                | Gr.                 | Top-k               | Nuc.                | Gr.                 | Top-k               | Nuc.                | Gr.                 | Top-k               |
| Pythia  | 2.8B      | 0.25                | 0.09                | 0.26                | 0.03                | 0.08                | 0.04                | 0.07                | 0.07                | 0.12                |
|         | 1B        | 0.21                | 0.12                | 0.27                | 0.02                | 0.04                | 0.04                | 0.08                | 0.04                | 0.18                |
| OLMo    | 2.1B      | 0.06                | 0.03                | 0.01                | 0.07                | 0.03                | 0.01                | 0.22                | 0.03                | 0.22                |
| Qwen3   | 0.6B      | 0.01                | 0.09                | 0.02                | 0.10                | 0.02                | 0.10                | 0.08                | 0.03                | 0.13                |
|         | 1.7B      | 0.02                | 0.08                | 0.03                | 0.05                | 0.01                | 0.03                | 0.09                | 0.04                | 0.14                |
|         | 8B        | 0.04                | 0.01                | 0.05                | 0.01                | 0.06                | 0.07                | 0.15                | 0.08                | 0.17                |
| Llama   | 3.2-3B    | 0.01                | 0.06                | 0.08                | 0.01                | 0.07                | 0.03                | 0.19                | 0.04                | 0.20                |
| Mistral | 7B-v0.3   | 0.00                | 0.01                | 0.06                | 0.04                | 0.02                | 0.01                | 0.16                | 0.14                | 0.12                |
| Mean    |           | 0.08 <sup>9/8</sup> | 0.06 <sup>6/8</sup> | 0.10 <sup>8/8</sup> | 0.04 <sup>4/8</sup> | 0.04 <sup>5/8</sup> | 0.04 <sup>4/8</sup> | 0.13 <sup>8/8</sup> | 0.06 <sup>5/8</sup> | 0.16 <sup>6/8</sup> |

Table 50: Drifting- and intermediate-feature Cohen’s  $d$  (gen-1 vs. gen-2+) under three decoding strategies per generator. Drifting features are expected to maintain large  $d$  regardless of strategy.

| Family  | Generator | entropy_mean        |                     |                     | TTR                 |                     |                     | hapax_ratio         |                     |                     | entropy_std         |                     |                     |
|---------|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|         |           | Nuc.                | Gr.                 | Top-k               | Nuc.                | Gr.                 | Top-k               | Nuc.                | Gr.                 | Top-k               | Nuc.                | Gr.                 | Top-k               |
| Pythia  | 2.8B      | 1.20                | 1.43                | 0.70                | 0.77                | 0.98                | 0.41                | 0.67                | 0.87                | 0.37                | 0.39                | 0.93                | 0.10                |
|         | 1B        | 1.40                | 0.98                | 0.83                | 0.80                | 0.64                | 0.41                | 0.65                | 0.53                | 0.34                | 0.53                | 0.91                | 0.13                |
| OLMo    | 2.1B      | 1.19                | 1.29                | 0.73                | 1.08                | 0.89                | 0.70                | 1.00                | 0.80                | 0.69                | 0.61                | 0.95                | 0.35                |
| Qwen3   | 0.6B      | 0.92                | 0.82                | 0.80                | 0.63                | 0.51                | 0.58                | 0.50                | 0.46                | 0.52                | 0.56                | 0.72                | 0.30                |
|         | 1.7B      | 1.12                | 1.03                | 0.99                | 0.93                | 0.74                | 0.85                | 0.83                | 0.68                | 0.80                | 0.52                | 0.78                | 0.32                |
|         | 8B        | 0.86                | 0.98                | 0.76                | 0.71                | 0.76                | 0.68                | 0.66                | 0.76                | 0.64                | 0.37                | 0.71                | 0.27                |
| Llama   | 3.2-3B    | 1.15                | 1.46                | 0.77                | 1.04                | 1.27                | 0.73                | 0.98                | 1.26                | 0.71                | 0.50                | 1.01                | 0.21                |
| Mistral | 7B-v0.3   | 0.89                | 1.31                | 0.70                | 0.82                | 1.04                | 0.64                | 0.82                | 1.03                | 0.67                | 0.23                | 0.62                | 0.15                |
| Mean    |           | 1.09 <sup>9/8</sup> | 1.16 <sup>6/8</sup> | 0.79 <sup>7/8</sup> | 0.85 <sup>7/8</sup> | 0.85 <sup>7/8</sup> | 0.63 <sup>5/8</sup> | 0.76 <sup>6/8</sup> | 0.80 <sup>7/8</sup> | 0.59 <sup>7/8</sup> | 0.46 <sup>5/8</sup> | 0.83 <sup>7/8</sup> | 0.23 <sup>5/8</sup> |

ity of autoregressive generation independent of sampling truncation. Top- $k$  sampling ( $k=50$ ) partially attenuates the partition. Shape features show boundary violations in 4/8 generators under top- $k$ ; however, these generators exhibit comparable or higher violations under nucleus sampling (Pythia entropy\_max: 0.21–0.25 nucleus vs. 0.26–0.27 top- $k$ ; OLMo tail\_decay: 0.22 under both). These are model-intrinsic boundary effects rather than strategy-induced degradation. Core drifting effect sizes weaken (mean  $d=0.67$  vs.  $0.90$  under

nucleus). This modulation is consistent with the mechanistic account in §5: top- $k$  truncation introduces a fixed-width selection window that interacts differently with depth-compressed distributions than nucleus sampling’s probability-mass truncation, partially blurring the location/shape boundary. The core DDR partition (drifting features carrying depth signal while shape features provide cross-generator reference) remains intact under both greedy and nucleus sampling and is attenuated but recognizable under top- $k$ . Entropy\_std exhibits strong strategy dependence (greedy mean  $d=0.83$ ; nucleus 0.46; top- $k$  0.23). Its depth sensitivity is modulated by the token-selection mechanism: deterministic greedy selection amplifies entropy variance across depths, while top- $k$ ’s fixed window compresses it. This further supports its classification as an intermediate rather than core feature.

## z.7 9-Generator Scale Evaluation

This analysis extends the 8-generator LOGO evaluation to nine generators by adding Gemma-2-2B, a model from a previously unrepresented architecture family (Gemma). All results use class-balanced LOGO; the 8-gen averages here (0.770/0.701) differ slightly from Table 2 (0.762/0.694), which uses an unfiltered protocol. Table 51 reports per-generator LOGO performance under this expanded set.

Table 51: 9-generator LOGO evaluation. Each row holds out the listed generator at test time. 3-class: gen-0/gen-1/gen-2+; binary: gen-1 vs. gen-2+. 95% bootstrap CIs in brackets.

| Generator       | Family  | Params | 3-class F1 | 95% CI         | Binary F1 | 95% CI         |
|-----------------|---------|--------|------------|----------------|-----------|----------------|
| Pythia-1B       | Pythia  | 1.0B   | 0.857      | [0.848, 0.866] | 0.785     | —              |
| Qwen3-0.6B      | Qwen    | 0.6B   | 0.819      | [0.807, 0.830] | 0.743     | —              |
| Pythia-2.8B     | Pythia  | 2.8B   | 0.812      | [0.801, 0.822] | 0.728     | —              |
| Qwen3-1.7B      | Qwen    | 1.7B   | 0.811      | [0.800, 0.823] | 0.746     | —              |
| OLMo-2-1B       | OLMo    | 1.5B   | 0.773      | [0.762, 0.784] | 0.692     | —              |
| Llama-3.2-3B    | Llama   | 3.2B   | 0.759      | [0.748, 0.770] | 0.680     | —              |
| Qwen3-8B        | Qwen    | 8.2B   | 0.696      | [0.683, 0.709] | 0.647     | [0.631, 0.664] |
| Gemma-2-2B      | Gemma   | 2.6B   | 0.628      | [0.622, 0.634] | 0.440     | [0.432, 0.449] |
| Mistral-7B      | Mistral | 7.0B   | 0.605      | [0.593, 0.617] | 0.534     | —              |
| Average (9-gen) |         |        | 0.751      |                | 0.666     |                |
| Average (8-gen) |         |        | 0.770      |                | 0.701     |                |

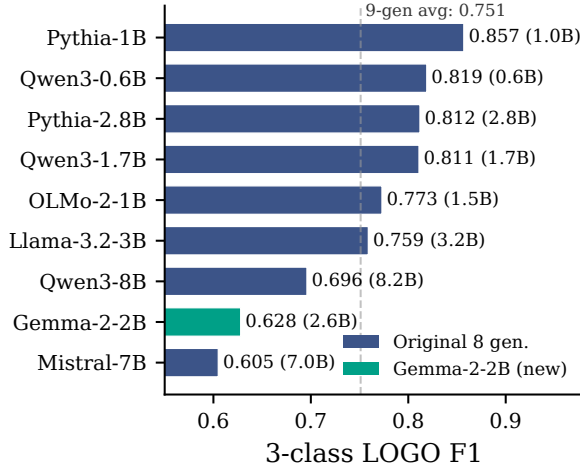


Figure 20: Per-generator 3-class LOGO F1 for the 9-generator evaluation. Gemma-2-2B (orange) is the weakest held-out generator. Dashed line: 9-gen average (0.751). Parameter counts shown in parentheses.

**Overall comparison.** Adding Gemma-2-2B reduces the mean 3-class LOGO F1 from 0.770 (8-gen) to 0.751 (9-gen), a drop of  $\Delta = -0.019$ . The per-generator F1 values for the original eight generators shift by at most  $[-0.011, +0.002]$ ; the inclusion of Gemma data in training does not destabilize existing generalization patterns.

**Gemma-2-2B anomaly.** Gemma-2-2B is the weakest held-out generator: 3-class F1 = 0.628, binary F1 = 0.440. Per-class breakdown reveals a striking asymmetry: gen-0 F1 = 0.991 and gen-2+ F1 = 0.856, but gen-1 F1 = 0.036; the classifier almost entirely fails to identify first-generation Gemma continuations. This is not a data-quality artifact: the Gemma-2-2B partition contains 4,957 instances with a balanced depth distribution ( $\sim 990$  per level). A plausible explanation is that Gemma-2’s first-generation continuations occupy a region of feature space highly similar to human-written text (gen-0), causing systematic misclassification of gen-1 as gen-0 or gen-2+. This constitutes

an architecture-specific generalization limitation rather than a systemic failure of the depth profiling approach. Several architectural differences may explain this behavior: (a) Gemma-2’s distinct attention mechanism and SentencePiece tokenizer with a larger vocabulary may produce first-generation continuations whose feature-space signature closely resembles human text, causing systematic gen-0/gen-1 confusion; (b) this anomaly does not affect DDR’s validity for the remaining eight generators, where per-generator F1 shifts by at most 0.011 upon Gemma inclusion.

**Scale effect.** A clear parameter-count gradient emerges across the nine generators: three of four sub-1.7B generators achieve  $F1 > 0.81$  (OLMo-2-1B at 0.773 is the exception), mid-range models (2–3B) cluster at 0.63–0.81, and large models (7B+) fall to 0.60–0.70. Qwen3-8B (8.2B parameters) achieves 3-class F1 = 0.696 under the 9-gen split, compared to 0.701 in the class-balanced 8-gen evaluation; this small difference reflects the changed training composition rather than a property of the Qwen3-8B generator itself.

**14B-parameter evidence.** Preliminary evidence from a 14.7B generator (Qwen3-14B-Base) confirms that the DDR partition persists at larger scale: Cohen’s  $d$  for drifting features remains in the large-effect range ( $d=0.78$ – $1.12$ ), though Spearman  $|\rho|$  attenuates by 21–35%, with entropy mean falling just below the drifting threshold ( $|\rho|=0.39 < 0.40$ ,  $d=0.78$ ). This scale-dependent signal compression is consistent with larger models producing more self-typical outputs that narrow inter-depth contrasts; the partition structure is preserved even as absolute effect sizes compress.

## z.8 Comparison with Concurrent Iterative-Generation Studies

Table 52 compares this work with two concurrent studies on iterative LLM generation. All three examine how repeated inference transforms text, but differ in chain operator, analytic lens, and downstream goal. Our continuation-chain setting permits richer depth signal and enables the location/shape decomposition that defines DDR.

**Finding-level connections.** Mohamed et al. (2025)’s factuality degradation under back-translation chains parallels DDR’s location drift: both document monotonic loss of input-specific information with iteration depth, though the for-

Table 52: Comparison with concurrent iterative-generation studies. All three examine how repeated LLM inference transforms text, but differ in chain operator, analytic lens, and downstream goal.

| Study                 | Chain Type                                                 | Feature Space                                                                              | Analysis Goal                                                           | Generator Scale             | Steps | Core Insight                                                                                                       |
|-----------------------|------------------------------------------------------------|--------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|-----------------------------|-------|--------------------------------------------------------------------------------------------------------------------|
| Mohamed et al. (2025) | Back-translation (6 bridge langs); rephrasing (ablation)   | Surface overlap (BLEU, ROUGE, METEOR, ChrF, BERTScore) + atomic factuality (FActScore)     | Factuality degradation rate across iterations                           | 7–9B (3 open models)        | 100   | Factual accuracy decays progressively; rate governed by source–bridge linguistic distance                          |
| Geng et al. (2026)    | Rephrasing + round-trip translation                        | String recurrence time, distinct-output count, similarity plateau (METEOR), KL contraction | Convergence & diversity dynamics of inference-time Markov chains        | 7–8B (3 open + GPT-4o-mini) | 50    | Greedy decoding collapses to 1–4 recurrent outputs; sampling sustains 10–20+ realizations                          |
| Ours                  | Open-ended continuation (no meaning-preserving constraint) | Token-level NLL statistics (location, scale, shape) + lexical diversity                    | Depth classification + DDR decomposition into generator & depth effects | 0.6–8B (8 models)           | 5     | NLL location drifts monotonically with depth while shape remains invariant, enabling per-generator depth profiling |

mer measures semantic content loss while DDR captures distributional compression in the NLL feature space. Their finding that degradation rate scales with source–bridge linguistic distance mirrors our observation that elaboration prompts attenuate DDR (Appendix c): tighter semantic constraints slow both factuality decay and distributional drift.

Geng et al. (2026)’s convergence dynamics under greedy decoding (collapse to 1–4 recurrent outputs within 50 steps) aligns with our greedy-decoding experiment where the DDR partition persists: shape features remain near-stationary (mean  $|d|=0.05$ ) while location drift is maintained (mean  $d=0.94$ ; Appendix z.6). This confirms that DDR is an empirical regularity of autoregressive generation independent of sampling stochasticity. Their KL contraction analysis complements DDR: where they quantify *how fast* the output distribution contracts, DDR decomposes *what* contracts (location) from what is preserved (shape).

## z.9 Token-Level NLL Distribution

Figure 21 plots the kernel density of per-token NLL across depths 0–4 (8 generators, 25 texts/depth,  $\sim 236k$  tokens, Pythia-410M reference). The distributions shift leftward as depth increases: median NLL drops from 1.64 nats (gen-0) to 0.02 (gen-4). Recursively generated text becomes increasingly predictable. This monotonic compression provides a mechanistic explanation for the entropy-based features driving the depth classifier.

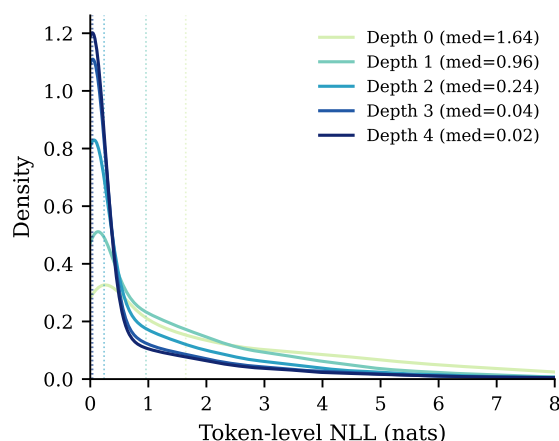


Figure 21: Token-level NLL distribution by generation depth, aggregated across all 8 generators (Pythia-410M reference model). Dashed lines indicate median NLL per depth. The systematic leftward shift demonstrates that deeper-generation text is increasingly predictable.

## aa Use of AI Assistants

We used Claude (Anthropic) as an AI coding and writing assistant throughout this work. Specifically:

- **Code:** drafting experiment scripts, data-processing pipelines, and analysis utilities; debugging and refactoring.
- **Writing:** drafting and revising paper sections, polishing prose, and reformatting  $\text{\LaTeX}$ .
- **Literature:** summarizing related papers to inform the related-work section.