The Truncation Blind Spot: How Decoding Strategies Systematically Exclude Human-Like Token Choices

Esteban Garcés Arias[1,2*] **Nurzhan Sapargali**[1]
Christian Heumann[1] **Matthias Aßenmacher**[1,2]

[1]Department of Statistics, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[*]Corresponding author: Esteban.GarcesArias@stat.uni-muenchen.de

Abstract

Why does machine-generated text remain detectable? We investigate a mechanistic explanation at the decoding stage: standard strategies such as top-k and nucleus sampling restrict generation to high-probability tokens, while human writers routinely choose contextually appropriate words from deeper in the model's probability distribution. Truncation makes a measurable share of these choices unreachable; we call this the *truncation blind spot*. Across five open models and three domains, 8–18% of human-selected tokens fall outside common truncation boundaries. Linguistic analysis further reveals disproportionate exclusion of content-word tokens. In a benchmark comprising 1.8 million machine generations, classifiers using only predictability and lexical diversity achieve mean AUC-ROC near 0.97, with substantial variation across decoding settings and strong transfer across generators. Probability-floor samplers substantially narrow the blind spot, demonstrating that the choice of truncation criterion matters for retaining human-used tokens. Together, these findings characterize a source of human–machine distributional mismatch and motivate decoding methods that preserve contextually appropriate low-probability choices while maintaining generation quality. Code and data are available at https://github.com/EstebanGarces/human_vs_machine.

1 Introduction

The ability to distinguish human-written text from machine-generated content has become increasingly important as large language models (LLMs) are widely deployed in applications ranging from content creation to automated communication (Jawahar et al., 2020; Tang et al., 2024). While a substantial body of work has focused on detection methods, a more fundamental question remains insufficiently understood: *why* is machine-generated text detectable in the first place? This paper investigates a mechanistic explanation grounded in the token selection of modern decoding strategies. Dominant approaches to open-ended generation, such as top-k sampling (Fan et al., 2018), nucleus sampling (Holtzman et al., 2020), and contrastive search (Su and Collier, 2023), share one property: they restrict generation to high-probability regions of the model distribution through likelihood-based truncation.

Human text, in contrast, routinely contains words that are contextually precise yet rank low under a model's next-token distribution. We measure this directly: the median rank of a human-selected token is 1–2, while the mean is 50–155 (Section 5.1). Exclusion arises from the interaction between the predictive distribution and the decoding rule. When many continuations have positive probability, some contextually acceptable choices can have moderate or low rank even under an accurate model. Truncation by rank or cumulative mass then cuts into this spread-out set, and the writer's specific choice often falls below the cutoff. Accounts of speech production describe how communicative goals guide lexical selection (Levelt, 1989), while relevance theory emphasizes the interaction between communicative intentions and context (Sperber and Wilson, 1986). We call the resulting tokens the **truncation blind spot**: tokens that human writers demonstrably use, that receive nonzero probability, and that standard decoding renders unreachable.

This mechanism offers one explanation for the reported convergence of different models toward similar outputs in open-ended settings (Jiang et al., 2025), since truncation constrains generation to similar regions of token space. We hypothesize that the same restriction induces systematic, detectable differences from human text, and investigate four questions: **RQ1** (magnitude): to what extent does likelihood-based truncation exclude tokens that humans actually select? **RQ2** (detection signal): does this mismatch provide a reliable sig-

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1302–1326
October 17-21, 2026 ©2026 Association for Computational Linguistics



From Wikinews: *"On 3 January 2026, the United States military invaded Venezuela, capturing and extrajudicially renditioning Venezuelan president Nicolás Maduro. [...] Maduro is to appear in a Manhattan federal court on 5 January 2026."*

Figure 1: Token ranks (log scale) of a human-written WikiNews passage. Red marks tokens excluded under top-k=10, with per-model exclusion rates. Top: GPT2-XL; bottom: Qwen2-7B. The passage postdates all evaluated models' training cutoffs (Section 5.1).

nal for distinguishing machine from human text? **RQ3** (mechanism primacy): is detectability governed by model characteristics or by the truncation mechanism? **RQ4** (quality trade-off): can configurations achieve low detectability while maintaining quality?

Since human token-selection distributions are not directly observable, we employ measurable proxies: *predictability* (average log-likelihood under a reference model) and *lexical diversity* (n-gram variety). We conduct a large-scale empirical study comprising 1,862,394 machine-generated texts from six open models, across five decoding strategies and 59 hyperparameter configurations, with supplementary proprietary and non-Transformer evaluations, alongside 5,261 human-written reference texts from three domains.

Contributions. Building on studies of decoding and detectability (Ippolito et al., 2020; Dubois et al., 2025) and theoretical analyses of truncation (Hewitt et al., 2022; Finlayson et al., 2024), we characterize the blind spot empirically and at scale: (i) a six-model detection grid with 59 decoding configurations and supplementary proprietary and non-Transformer evaluations; (ii) a part-of-speech and domain decomposition of the excluded tokens (Section 5.3) and a within-family analysis of truncation-set overlap across scale (Appendix A.24); (iii) comparisons across decoding settings, model scales, architectures, and proprietary systems, including cross-generator transfer with the generating model unseen at training time (Sections 5.4 and 5.5); and (iv) an evaluation of tail-recovering samplers against this benchmark: probability-floor criteria (ϵ, η) yield lower exclusion than the nucleus baseline in all five open models at the least restrictive evaluated settings while typicality-based truncation does not; generation experiments test the exclusion–detectability link out of sample (Sections 5.1, 5.2). Together these characterize a decoding-related source of distributional mismatch, giving detection research a mechanism to investigate and decoding research a measurable cost to assess.

2 Formal Setup: Truncation Sets and the Exclusion Rate

We define the objects our measurements operate on: the truncation set induced by a decoding strategy, the blind spot, and the exclusion rate that quantifies it. The quantity estimated in this formal setup is the exclusion rate of Equation 1, a sample statistic over observed human text that requires no model of the human production process.

2.1 Likelihood-Based Truncation in Language Models

Let $\mathcal{M}$ be an autoregressive language model with vocabulary $\mathcal{V}$. For each token $x_t \in \mathcal{V}$ at step t given context $\mathbf{x}_{<t} = (x_1, \ldots, x_{t-1})$, the model produces a categorical distribution $\mathbf{p}_t$ on the simplex $\Delta^{|\mathcal{V}|-1}$ with entries $P_{\mathcal{M}}(x_t \mid \mathbf{x}_{<t})$. For a truncation sampler, a *support-selection rule* $\mathcal{S} : \Delta^{|\mathcal{V}|-1} \to 2^{\mathcal{V}} \setminus \{\emptyset\}$ maps $\mathbf{p}_t$ to a truncation set $\mathcal{T} \subseteq \mathcal{V}$, from which the next token is sampled (renormalized). Top-k sampling retains the k most probable tokens, $\mathcal{T}(\mathbf{x}_{<t}) = \{x \mid \mathrm{rank}(x \mid \mathbf{x}_{<t}) \leq k\}$, where $\mathrm{rank}(x \mid \mathbf{x}_{<t}) = |\{v \in \mathcal{V} : P_{\mathcal{M}}(v \mid \mathbf{x}_{<t}) \geq P_{\mathcal{M}}(x \mid \mathbf{x}_{<t})\}|$, so that tied tokens share the rank of the last member of their tie group; nucleus (top-p) sampling retains the smallest top-ranked set with cumulative mass at least p; contrastive search uses a top-k candidate pool but selects by a score that also penalizes degeneration. Beam search prunes competing prefixes by sequence score; it is included in our empirical comparisons but is not represented by this token-local sampling rule. We define a neural text generator as $\theta = (\mathcal{M}, \mathcal{S})$ with truncation set $\mathcal{T}_\theta(\mathbf{x}_{<t})$ and *excluded region* $\mathcal{E}_\theta(\mathbf{x}_{<t}) = \mathcal{V} \setminus \mathcal{T}_\theta(\mathbf{x}_{<t})$, i.e., the tokens unreachable under θ.

2.2 The Blind Spot

One construct on the human side suffices: the *human truncation set* $\mathcal{T}_{\mathcal{H}}(\mathbf{x}_{<t}) \subseteq \mathcal{V}$, the tokens a writer would accept as continuations at position t. It is a definitional device rather than an estimand: we never measure $\mathcal{T}_{\mathcal{H}}$ itself, only the event that a token a writer actually produced, by construction an element of $\mathcal{T}_{\mathcal{H}}$, falls outside the model's truncation set. Our analysis concerns the *blind spot*, $\mathcal{T}_{\mathcal{H}} \cap \mathcal{E}_\theta$: tokens a writer selects that the model excludes. Appendix A.8 gives the full four-case taxonomy.

Prediction and reachability. Even an accurate predictive distribution can assign positive probability to more acceptable continuations than a truncation rule retains. For example, if $m > k$ acceptable tokens have positive probability, a top-k rule necessarily excludes some of them; a nucleus rule does so whenever its retained set omits an acceptable token. The blind spot is a property of the pair $\theta = (\mathcal{M}, \mathcal{S})$: improving the model can change which tokens are excluded, while changing the decoding rule can change their reachability without changing the model.

2.3 Quantifying the Blind Spot

We quantify how often human-selected tokens fall into the blind spot. Since every observed human token lies in $\mathcal{T}_{\mathcal{H}}$ by construction, the blind-spot event coincides with exclusion by the model:

$$x_t^H \in \mathcal{T}_{\mathcal{H}}(\mathbf{x}_{<t}^H) \cap \mathcal{E}_\theta(\mathbf{x}_{<t}^H) \iff x_t^H \notin \mathcal{T}_\theta(\mathbf{x}_{<t}^H).$$

Its magnitude is the rate at which human tokens fall outside the model's truncation set,

$$\mathcal{R}_{\text{excl}}(\mathbf{x}^H; \theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[x_t^H \notin \mathcal{T}_\theta(\mathbf{x}_{<t}^H)\right], \quad (1)$$

where $\mathbf{1}[\cdot]$ is the indicator function: a direct sample statistic that counts how often a writer's token would have been unreachable under generator θ, requiring no assumptions about the process that produced the text.

2.4 Measurable Proxies

Two text properties reflect the mismatch. *Predictability*: machine text constrained to high-probability tokens should show higher average log-likelihood under a reference model than human text. *Lexical diversity*: by concentrating mass on frequent tokens, truncation should reduce n-gram variety relative to human text, which draws on a broader effective vocabulary.

3 Related Work

Detection research spans supervised classifiers (Jawahar et al., 2020; Tang et al., 2024) and zero-shot statistical methods (Mitchell et al., 2023). Gehrmann et al. (2019) visualized that machine tokens concentrate in high-probability regions; we quantify the converse, how often human tokens fall outside truncation sets. Closest to our work, Ippolito et al. (2020) showed that decoding choices affect detectability, and Dubois et al. (2025) report that sampling parameters move detector accuracy.

We add the token-level mechanism behind these observations: a measured exclusion rate with a known linguistic composition, evidence relating this mechanism to detectability across models, and an evaluation of proposed remedies against it.

On the decoding side, truncation samplers trade quality against diversity (Fan et al., 2018; Holtzman et al., 2020; Su and Collier, 2023; Garces Arias et al., 2025c). Hewitt et al. (2022) cast truncation as desmoothing and propose η- and ϵ-sampling; Meister et al. (2023) select tokens by typicality. We take the complementary, diagnostic view, asking what truncation excludes relative to human text, and we evaluate these samplers against our exclusion benchmark (Sections 5.1 and 5.2). Adaptive Contrastive Search adjusts its degeneration penalty using uncertainty (Garces Arias et al., 2024), and GUARD combines local and global entropy signals (Ding et al., 2025). Min-k selects truncation boundaries from relative logit changes (Ding et al., 2026b). These methods offer further candidates for evaluating human-token retention. Psycholinguistic and information-theoretic accounts of language (Levelt, 1989; Jaeger, 2010) provide a complementary perspective on human token choice (extended discussion in Appendix A.7).

4 Experimental Setup

4.1 Language Models

We evaluate eight language models spanning sizes, architectures, and training procedures: the open-source Llama-3.1 (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen 2 (Yang et al., 2024), Falcon 2 (Malartic et al., 2024), Deepseek (Bi et al., 2024), and GPT2-XL (Radford et al., 2019), and the proprietary GPT-3.5-turbo and Claude-3-Haiku. We additionally include the non-Transformer architectures Mamba (Gu and Dao, 2024) and RWKV (Peng et al., 2023) to test whether the phenomenon is Transformer-specific. Proprietary systems are queried under provider API defaults, the conditions under which they are deployed; the exclusion analyses use the checkpoints listed in Appendix A.1.

4.2 Decoding Strategies and Configurations

Following Garces Arias et al. (2025c), we evaluate five decoding strategies: beam search (widths $\{3, 5, 10, 15, 20, 50\}$), contrastive search (Su and Collier, 2023) ($k \in \{1, 3, 5, 10, 15, 20, 50\}$, $\alpha \in \{0.2, \ldots, 1.0\}$), top-$k$ sampling (Fan et al., 2018) ($k \in \{1, 3, 5, 10, 15, 20, 50\}$), top-$p$ sampling (Holtzman et al., 2020) ($p \in \{0.6, 0.7, 0.8, 0.9, 0.95\}$), and temperature sampling ($T \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$), giving 59 configurations per model and 1,862,394 generated texts across the six open models. The four non-temperature strategies contribute 53 configurations. Temperature rescales probabilities without itself removing support; inherited sampling defaults can impose additional truncation. In Section 5.1 we additionally evaluate three tail-recovering truncation rules as *exclusion criteria* on human text: ϵ-sampling ($\epsilon \in \{3, 9, 20\} \times 10^{-4}$) and η-sampling ($\epsilon \in \{0.9, 2, 4\} \times 10^{-3}$; both rules are parameterized by ϵ in Hewitt et al., 2022), and locally typical sampling ($\tau \in \{0.2, 0.9, 0.95\}$) (Meister et al., 2023); Section 5.2 additionally generates with a subset of these rules for two models. Grid generation follows the generation library's default settings, under which the truncation applied in total can be stricter than the nominal parameter alone; the harness of Section 5.2 instead applies each rule in isolation.

4.3 Datasets

We employ three datasets representing distinct writing styles, totaling 5,261 human-written texts: BookCorpus (Zhu et al., 2015) (1,947 samples of creative fiction), WikiText (Merity et al., 2016) (1,314 Wikipedia articles), and WikiNews[1] (2,000 news reports).

4.4 Metrics

Predictability. Given a text $\mathbf{x} = (x_1, \ldots, x_T)$, prompt $\mathbf{c}$, and reference model $\mathcal{M}_{\text{ref}}$:

$$\text{Pred}(\mathbf{x} \mid \mathbf{c}) = \frac{1}{T} \sum_{t=1}^{T} \log P_{\mathcal{M}_{\text{ref}}}(x_t \mid [\mathbf{c} : \mathbf{x}_{<t}]). \tag{2}$$

Higher (less negative) values indicate higher likelihood; predictability is the negative of mean per-token log-perplexity, and we use the term for accessibility. We use OPT-2.7B (Zhang et al., 2022) as $\mathcal{M}_{\text{ref}}$, following the evaluation setup of Garces Arias et al. (2025a); results are robust to the choice of judge model (pairwise $\rho \geq 0.88$; Appendix A.20).

Lexical Diversity. We measure n-gram diversity as

$$\text{Div}(\mathbf{x}) = 100 \cdot \prod_{n=2}^{4} \frac{|\text{unique } n\text{-grams}(\mathbf{x})|}{|\text{total } n\text{-grams}(\mathbf{x})|}. \tag{3}$$

[1] http://www.wikinews.org

Setting	Book	News	Wiki	Avg. (95% CI)
$k = 5$	25.0	22.5	27.4	25.0 [23.1, 27.4]
$k = 10$	18.1	15.7	20.4	18.1 [16.3, 20.2]
$k = 20$	12.9	10.8	15.2	13.0 [11.4, 14.8]
$k = 50$	8.0	6.5	9.9	8.1 [6.9, 9.5]
$p = 0.6$	28.1	28.5	29.6	28.7 [27.7, 29.8]
$p = 0.9$	7.4	8.7	7.9	8.0 [7.4, 8.8]
$p = 0.95$	3.5	4.5	3.7	3.9 [3.4, 4.6]

Table 1: Human token exclusion rates $\mathcal{R}_{\text{excl}}$ (%) at representative top-k and top-p settings, averaged over five open models (GPT2-XL, Qwen2-7B, Mistral-7B-v0.3, Llama-3-8B, Falcon2-11B) and computed within each model's own tokenizer. Full sweep in Table 9 (Appendix A.11).

Low diversity indicates repetition; high diversity can reflect genuine richness or incoherent randomness. The two features enter all analyses separately: low entropy at generation time need not imply low lexical diversity, and the two properties capture complementary aspects of the mismatch.

4.5 Classification Setup

We train binary classifiers, human-written versus machine-generated text from a given (model, configuration) cell, on the two features alone: Random Forest (Breiman, 2001) (RF, 500 trees), Logistic Regression (LR), and Naive Bayes (NB), with 80-20 class-stratified splits of individual texts for the per-configuration grid. Human and machine continuations from the same prompt can occur in different partitions. Simple classifiers are intentional: high performance with two features indicates systematically different statistics. AUC-ROC is invariant to the class prior, and downsampling machine examples leaves it essentially unchanged (Appendix A.14).

5 Results

5.1 The Magnitude of the Blind Spot (RQ1)

Figure 1 illustrates the blind spot and Table 1 quantifies it: at k=10, roughly 18% of human-selected tokens are unreachable; at k=50, roughly 8%; nucleus sampling at p=0.9 excludes about 8%. Exclusions concentrate on content words in our separate GPT-2 analysis (Section 5.3); token-level rates alone do not determine passage-level reachability because exclusions can be dependent. The full sweep and rank distributions are in Appendix A.11. This is the signature of the mechanism in Section 2.2: at most positions the writer's token is where the model expects it, but at positions of lexical freedom the model's mass is spread over many plausible candidates and fixed cutoffs remove the chosen one. WikiText shows the heaviest tail, consistent with specialized vocabulary. The Figure 1 passage postdates every model's training cutoff, and its exclusion rates (24.6%/14.8% at k=10) are comparable to the WikiNews average (15.7%), showing that exclusion also occurs in text written after pretraining.

Tail-recovering truncation criteria. Our account makes a testable prediction: truncation criteria that impose no fixed rank or mass budget should retain more of the human-used tail at matched permissiveness. We evaluate the exclusion side of this prediction for η- and ϵ-sampling (Hewitt et al., 2022) and locally typical sampling (Meister et al., 2023) on the same human corpora, for the five open models of Table 1 (Table 2). Because average truncation-set sizes are tokenizer-dependent, we compare within each model.

Probability-floor criteria confirm the prediction, and more strongly than matched permissiveness requires. In every model, both ϵ- and η-sampling admit a setting that strictly dominates nucleus sampling at p=0.9, with lower exclusion at a smaller average truncation set and relative reductions of 23–38% across all ten (model, rule) pairings (e.g., Falcon2-11B: ϵ-sampling at 6.2% exclusion and 39 tokens vs. nucleus at 10.1% and 53). Aggregated over models and domains, exclusion falls from 8.8% (nucleus, p=0.9) to 5.3% (η) and 4.2% (ϵ) at comparable retained mass (Table 2). Throughout the table, exclusion closely tracks the truncated probability mass ($\mathcal{R}_{\text{excl}} \approx 100 -$ retained mass): human choices occupy the model's mass in proportion to it, an aggregate correspondence that does not, by itself, establish model calibration or rule out model error (Section 2.2).

Locally typical sampling, by contrast, tracks nucleus sampling within 0.1 points in every model at matched mass (τ=0.9: 8.8% for both; τ=0.95: 4.7% for both) and at low τ *amplifies* the blind spot (τ=0.2: 54.9%), since the human-used tail consists precisely of the high-surprisal tokens that typicality treats as atypical. The operative distinction is therefore not likelihood-based versus information-based selection, but whether the criterion caps retained mass or rank (nucleus, top-k, typicality) or sets a per-token probability floor (ϵ, η) that widens the pool where the distribution is flat, exactly the

Rule	Param	$\mathcal{R}_{\text{excl}}$ (%)	$\overline{\lvert\mathcal{T}\rvert}$	Retained mass
top-k	$k = 10$	18.8	10	0.82
top-k	$k = 50$	8.1	50	0.92
top-p	$p = 0.6$	29.6	7	0.72
top-p	$p = 0.9$	8.8	86	0.92
top-p	$p = 0.95$	4.7	211	0.96
ϵ	3×10^{-4}	4.2	91	0.96
ϵ	9×10^{-4}	6.7	44	0.94
η	9×10^{-4}	5.3	89	0.95
η	2×10^{-3}	6.8	66	0.94
typical	$\tau = 0.2$	54.9	8	0.46
typical	$\tau = 0.9$	8.8	91	0.92
typical	$\tau = 0.95$	4.7	212	0.96

Table 2: Human-token exclusion under tail-recovering rules versus rank and mass baselines, averaged over the five open models of Table 1 (checkpoints in Appendix A.1); computed within each model's tokenizer, conditioned on the prompt, ties per Section 2. The k=10 baseline is 18.8% in this analysis and 18.1% in Table 1. Set sizes are tokenizer-dependent; dominance comparisons are made within each model (see text).

positions of greatest lexical freedom identified in Section 2.2. Even the best criterion leaves 4–5% of human tokens unreachable: probability-floor truncation narrows the blind spot substantially but does not close it. Removing truncation restores support but does not, by itself, ensure generation quality. Section 5.2 tests this prediction directly by generating with these rules.

5.2 From Exclusion to Detectability: An Out-of-Sample Test

The preceding analysis measures what these rules exclude; we now test whether the exclusion ordering predicts detectability for rules the detection grid has never seen. For GPT2-XL and Qwen2-7B we generate continuations of all 5,261 prompts under the four floor settings of Table 2, under locally typical sampling (τ=0.9), and under nucleus sampling (p=0.9) as baseline, each rule applied in isolation, scoring each configuration with the classifiers of Section 4.5. We evaluate two predictions: (P1) AUC follows the exclusion ordering of Table 2, placing lower-exclusion criteria below nucleus sampling; (P2) typical sampling at matched mass is indistinguishable from nucleus sampling.

Figure 2 reports each configuration's ΔAUC against its model's nucleus baseline (Table 4, Appendix A.3). Qwen2-7B supports P1: exclusion and AUC have perfect rank agreement across the evaluated configurations (Spearman $\rho = 1.00$), and the strongest floor setting reduces AUC by 0.14

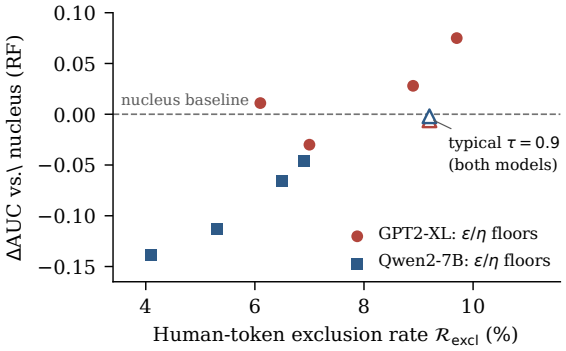


Figure 2: Exclusion and detectability for unseen sampling rules. Change in RF AUC-ROC relative to each model's nucleus baseline (dashed line at Δ=0) for each out-of-sample configuration (Table 4), against its human-token exclusion rate. Filled: probability-floor settings; open: typical sampling, close to the baseline.

(RF) to 0.16 (LR). GPT2-XL provides mixed support: rank agreement varies by classifier ($\rho = 0.37$–0.83), and some lower-exclusion settings increase AUC. For example, $\epsilon = 3 \times 10^{-4}$ lowers exclusion from 9.2% to 6.1% but increases RF AUC by 0.011. P1 is therefore not consistently confirmed for GPT2-XL. For P2, typical sampling differs from nucleus sampling by at most 0.006 AUC in both models and the reported RF/LR comparisons. These small observed differences are consistent with P2.

These results provide model-dependent support for a relationship between exclusion and detectability. Probability-floor criteria substantially reduce detectability for Qwen2-7B, while GPT2-XL shows a mixed response. Its scope is two models, two detection features, and no quality evaluation of floor-generated text; Hewitt et al. (2022) report separate human evaluations favoring η-sampling at comparable settings.

5.3 Linguistic Composition of the Blind Spot

We examine which tokens are excluded in a within-family analysis of GPT-2 small, medium, and large. A fixed sample contains 200 unique prompt/reference pairs per domain (600 total) from the benchmark's available reference checkpoints. We tag each prompt and continuation with spaCy (Honnibal et al., 2020) and align model subwords to Universal POS tags (Nivre et al., 2016) using character offsets. Exclusion is measured on up to 256 continuation subwords, conditioned on the prompt. The shared GPT-2 tokenizer permits comparisons across model sizes without changing subword units.

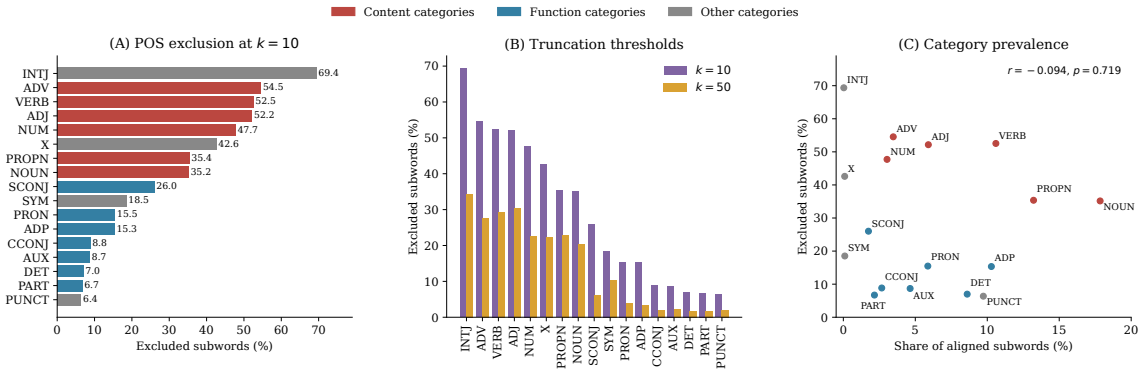


Figure 3: POS analysis of continuation subwords from 600 prompt/reference pairs, pooled across three GPT-2 checkpoints. **(A)** Exclusion at $k = 10$: content categories (red), function categories (blue), and other tags (gray). **(B)** Exclusion at $k = 10$ (purple) and $k = 50$ (gold, hatched). **(C)** POS-category prevalence versus exclusion ($r = -0.09$). Rates count aligned subwords, not whole words.

Appendix A.26 details sampling, alignment, and aggregation.

At $k = 10$, pooled exclusion is 42.4% for subwords aligned to content words and 12.0% for those aligned to function words: a ratio of 3.53 (95% stratified prompt/reference-pair bootstrap CI: 3.43–3.64). The content–function ordering holds in all nine model/domain combinations (ratios 3.27–3.84). Category-specific rates range from 35.2–54.5% for content categories and 6.7–26.0% for function categories (Figure 3A). At $k = 50$, both pooled rates decrease, to 24.4% and 2.7%, respectively.

5.4 Detection Signals Across Models and Decoding Settings (RQ2, RQ3)

Using only predictability and diversity, all three classifiers achieve mean AUC-ROC of 0.968–0.974 and mean F1 of 0.949–0.953 across the 1,062 (model, domain, decoding configuration) cells of the grid (Table 11, Appendix A.12). Strong performance across these classifiers shows that the two features carry substantial detection signal.

Detectability varies with both the evaluated models and their decoding settings (Appendix A.12). Llama-3.1 has the highest mean RF AUC (0.993), while Falcon 2 has the lowest (0.945); these rankings do not follow model age or size alone. However, inherited generation defaults confound a comparison of model identity at a nominally fixed decoding setting (Section 4.2). Within-family scale coefficients are non-significant for three of four families, but GPT-2 shows a significant negative effect ($\beta = -0.052$, $p = 0.015$), as does the pooled estimate ($\beta = -0.021$, $p = 0.017$). Non-Transformer systems have higher detectability in the architecture analysis (ΔAUC $= +0.180$). High detectability therefore persists across the evaluated architectures, with measurable variation by model characteristics.

A within-family analysis of the truncation sets points the same way: models of one family share substantially overlapping sets (mean Jaccard at k=10: 0.589 for GPT-2, 0.531 for Qwen2), and overlap decreases only gradually with the size gap. These overlaps quantify shared candidates across model sizes (Appendix A.24). Truncation parameters, by contrast, show a clear relationship with detectability: beam search is most detectable (AUC-ROC 0.997), followed by contrastive search (0.981), top-p (0.956), and top-k (0.948), a 4.9-point spread among these strategy averages (Table 3, Appendix A.2). The strategy averages summarize different parameter grids and include all effects of each decoding configuration. For top-k, increasing k from 1 to 15 reduces AUC-ROC from 0.988 to 0.924; for nucleus sampling, increasing p from 0.6 to 0.95 reduces AUC-ROC from 0.978 to 0.935. Even the least detectable configurations do not fully match human statistics (human baseline: predictability -2.73, diversity 94.11); permissive truncation reduces but does not eliminate the mismatch (Appendix A.18).

5.5 Cross-Model Generalization

So far, classifiers were trained per (model, strategy, configuration) cell. We now ask whether detection generalizes when the configuration is unknown and the generating model is unseen at training time, using the six open models. Classifiers here train on

data pooled across all 53 configurations per model, so the same test set can be reused across training regimes; pooling adds within-class variance, so absolute AUC is lower (in-distribution baseline 0.950 vs. per-cell 0.974); the relevant statistic is the difference between matched conditions on the same test set.

Matched training-set comparison. For each model M we evaluate three conditions on the same held-out test set: **ID-single** trains on M alone; **ID-pooled** trains on all six models including M; and **LOMO** trains on all models except M, so M is unseen. ID-pooled and LOMO are matched in training-set size, so $\Delta = \text{AUC}_{\text{ID-pooled}} - \text{AUC}_{\text{LOMO}}$ isolates the contribution of M's identity in training.

Across the six held-out models, leave-one-model-out training yields mean RF AUC-ROC of 0.929, and Δ is negative for every model (mean -0.0044, 95% CI $[-0.0063, -0.0025]$; paired $t(5) = -5.90$, $p = 0.002$); for NB and LR, Δ is indistinguishable from zero (Appendix A.13, Table 6). Including the held-out model does not improve mean RF performance in this comparison; the NB and LR intervals include zero. Transfer indicates shared feature-level regularities across generators, consistent with a decoding-related distributional mismatch. Identifying the relative contributions of selection, exclusion, and repetition to this transfer requires further analysis.

GPT-2 $\rightarrow$ modern transfer. A complementary test asks whether the signal learned from the oldest, smallest model transfers to far more capable ones. We trained a classifier only on GPT2-XL outputs and evaluated it on each modern open model, alongside a target-specific in-distribution baseline on the same held-out test set.

The classifier trained only on GPT2-XL achieves mean RF AUC 0.949 across the modern targets, within 0.002 of the target-specific baselines (mean $\Delta = +0.0015$, 95% CI $[-0.0022, +0.0051]$; Appendix A.13); on Llama it reaches 0.987 against an in-distribution 0.991. A classifier trained only on 2019-era outputs thus detects substantially more capable models at near in-distribution accuracy, consistent with the structural account.

Proprietary systems. GPT-3.5-turbo and Claude-3-Haiku, both proprietary and extensively aligned, remain highly detectable, with AUC-ROC above 0.98 for all three classifiers (Appendix A.17).

5.6 The Detectability–Quality Dissociation (RQ4)

The two detection features do not measure contextual appropriateness: high diversity can occur in incoherent text as well as in human writing. Table 7 (Appendix A.6) illustrates repetitive and incoherent machine continuations. Both examples receive high estimated machine probabilities.

We quantify the dissociation with public human ratings from Garces Arias et al. (2025a) (Likert 1–5, $n = 300$). The correlation between quality and estimated machine probability is small and non-significant (Spearman $\rho = 0.097$, $p = 0.092$). Beam search, the most detectable strategy, receives the lowest mean rating (2.00), while contrastive search, whose configurations include the least detectable in our grid (AUC-ROC as low as 0.53; Appendix A.16), is rated 2.30 against a human reference of 3.26 (Figure 9, Appendix A.25).

The detectors we study are therefore best understood as detectors of likelihood-constrained generation, not of machine text in general; low classifier confidence does not imply improved quality. Practical recommendations are collected in Appendix A.10.

6 Conclusion

Likelihood-based truncation creates a blind spot between the continuations a model can represent and those its decoder can produce. Our measurements show that human writers regularly select tokens outside common truncation boundaries, with disproportionate exclusion of content-word tokens. Across a large generation benchmark, predictability and lexical diversity provide strong detection signals that transfer across generators, while decoding choices substantially affect detectability. Probability-floor criteria recover more of the human-used tail, pointing to opportunities for improving token selection. These findings motivate evaluating next-token prediction and decoding together: a model can assign positive probability to a human choice while its decoder makes that choice unreachable. The exclusion rate provides a concrete diagnostic for this loss, complementing evaluations of generation quality. Preserving human-like variation requires retaining contextually appropriate choices without sacrificing coherence; lower detectability alone is not sufficient evidence of better generation. Addressing this challenge is an important direction for decoding research.

Limitations

(1) Predictability and diversity proxy distributional mismatch but not contextual appropriateness; text can achieve human-like statistics while remaining incoherent. (2) We observe only what writers produced, not why; claims about the communicative function of excluded tokens are interpretive. The POS analysis covers a separate 600-pair subset and three GPT-2 checkpoints, uses automatic tags and subword units, and excludes ambiguous alignments. Its category-prevalence analysis does not control for individual word frequency. Reference-model agreement and within-tokenizer measurement provide robustness checks but do not rule out all modeling or tokenization effects. (3) Scope is constrained to English text from three domains (fiction, encyclopedia, news) at $\leq$ 256 tokens; patterns may differ across languages, genres, or longer contexts, and the binary distinction simplifies post-edited scenarios. Since BookCorpus and WikiText likely appear in pretraining corpora, the 8–18% exclusion may understate the true magnitude; the post-cutoff WikiNews passage in Figure 1 exhibits comparable exclusion (Section 5.1), but one passage does not establish a lower bound for unseen text. (4) Exclusion analyses cover the five open-source models of Table 1; the tail-criterion detectability test (Section 5.2) covers two models and our two features, without quality evaluation; results may vary with other architectures. (5) The per-configuration detection grid splits individual texts, so its test performance does not establish generalization to unseen prompts. (6) For proprietary systems, decoding parameters cannot be controlled beyond API defaults, so the effect of alignment is not identified. The main detection grid also uses inherited generation defaults; nominal settings do not isolate a single truncation rule.

These limitations motivate the natural next steps: extending the out-of-sample detectability experiments of Section 5.2 to all models and strategies with human quality evaluation; benchmarking further adaptive rules such as min-p (Nguyen et al., 2025), Mirostat (Basu et al., 2021), and adaptive contrastive search (Garces Arias et al., 2024) on the exclusion diagnostic; and developing adaptive truncation that relaxes bounds at content-bearing positions. Exclusion measures the reachability of human choices, not generation quality, and rules may rank differently under quality-based criteria.

Ethics Statement

We use publicly available text datasets, including encyclopedic and news material that can mention identifiable people. Understanding detectability could inform attempts to evade detection; we judge the scientific value of understanding why machine text is detectable to outweigh this risk, particularly as our findings point toward improving generation rather than evading detection. Code and supporting data are available at `https://github.com/EstebanGarces/human_vs_machine`. We declare no conflicts of interest.

Acknowledgments

Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI, under Grant No.460037581. Esteban Garcés Arias and Matthias Aßenmacher acknowledge the Munich Center for Machine Learning (MCML) for providing financial support and computational resources that enabled the experimental component of this work.

References

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Yuanhao Ding, Esteban Garces Arias, Meimingwei Li, Julian Rodemann, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, and Chongsheng Zhang. 2025. GUARD: Glocal uncertainty-aware robust decoding for effective and efficient open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7202–7226, Suzhou, China. Association for Computational Linguistics.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. 2026a. Breaking the likelihood trap: Variance-calibrated modulation for large language model decoding. *Preprint*, arXiv:2606.22511.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. 2026b. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14932–14948, San Diego, California, United States. Association for Computational Linguistics.

Matthieu Dubois, François Yvon, and Pablo Piantanida. 2025. How sampling affects the detectability of machine-written texts: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11369–11387, Suzhou, China. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. 2024. Closing the curious case of neural text degeneration. In *The Twelfth International Conference on Learning Representations*.

Giacomo Frisoni, Luca Ragazzi, David Cohen, Gianluca Moro, Antonella Carbonaro, and Claudio Sartori. 2026. Abstractive summarization through the prism of decoding strategies. *Neural Networks*, 196:108249.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Matthias Assenmacher, and Christoph Jansen. 2025a. Statistical multicriteria evaluation of LLM-generated text. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 338–351, Hanoi, Vietnam. Association for Computational Linguistics.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025b. Towards better open-ended text generation: A multicriteria evaluation framework. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025c. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.

Esteban Garces Arias, Julian Rodemann, and Christian Heumann. 2025d. The geometry of creative variability: How credal sets expose calibration gaps in language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 36–50, Suzhou, China. Association for Computational Linguistics.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2024. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Edward Gibson, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Aaron Grattafiori, Abhimanyu Dubey, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Peter Indefrey and Willem JM Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Preprint*, arXiv:2510.22954.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36.

Willem JM Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, Mohammed Al-Yafeai, Hamza Alobeidli, Leen Al Qadi, Mohamed El Amine Seddik, Kirill Fedyanin, Reda Alami, and Hakim Hacid. 2024. Falcon2-11B technical report. *arXiv preprint arXiv:2407.14885*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 24950–24962. PMLR.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, and 13 others. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell, Oxford.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation. *Transactions on Machine Learning Research*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting LLM-generated text. *Communications of the ACM*, 67(4):50–59.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators.

Transactions of the Association for Computational Linguistics, 10:997–1012.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Preprint*, arXiv:1506.06724.

A Appendix

A.1 Model Checkpoints

The five open models used in both exclusion analyses (Tables 1 and 2) correspond to the checkpoints `gpt2-xl`, `Qwen/Qwen2-7B`, `mistralai/Mistral-7B-v0.3`, `meta-llama/Meta-Llama-3-8B`, and `tiiuae/falcon-11B`.

A.2 Detection by Decoding Strategy

Table 3 reports detection performance by decoding strategy, referenced from Section 5.4.

Strategy	RF		LR		NB	
	AUC	F1	AUC	F1	AUC	F1
Beam	0.997	0.991	0.996	0.990	0.997	0.989
CS	0.981	0.962	0.970	0.954	0.977	0.954
Temperature	0.957	0.934	0.958	0.938	0.959	0.938
Top-p	0.956	0.918	0.958	0.921	0.957	0.919
Top-k	0.948	0.913	0.951	0.918	0.951	0.917

Table 3: Detection performance by decoding strategy, ordered by mean RF AUC. Means weight model/domain/configuration cells equally within each strategy.

A.3 Out-of-Sample Detectability

Table 4 lists the per-configuration ΔAUC values underlying Figure 2 and the analysis of Section 5.2.

Rule, parameter	$\mathcal{R}_{\text{excl}}$ (%)	ΔAUC (RF/LR)
GPT2-XL		
ϵ, 3×10^{-4}	6.1	$+0.011/-0.105$
η, 9×10^{-4}	7.0	$-0.030/-0.061$
ϵ, 9×10^{-4}	9.7	$+0.075/+0.078$
η, 2×10^{-3}	8.9	$+0.028/+0.026$
typical, τ=0.9	9.2	$-0.006/-0.006$
nucleus, p=0.9	9.2	0/ 0
Qwen2-7B		
ϵ, 3×10^{-4}	4.1	$-0.139/-0.163$
η, 9×10^{-4}	5.3	$-0.113/-0.103$
ϵ, 9×10^{-4}	6.5	$-0.066/-0.051$
η, 2×10^{-3}	6.9	$-0.046/-0.039$
typical, τ=0.9	9.2	$-0.002/-0.003$
nucleus, p=0.9	9.2	0/ 0

Table 4: Out-of-sample detectability of tail-recovering rules: change in AUC-ROC relative to each model's within-harness nucleus baseline (RF/LR per cell; NB parallels these), prompt-level splits, pooled over domains, each rule applied in isolation; $\mathcal{R}_{\text{excl}}$ is the per-model exclusion rate. Negative values indicate reduced detectability. Human reference features in this experiment are Pred -2.70 and Div 94.18, compared with -2.73 and 94.11 in the detection grid (Table 21). The nominal p=0.9 exclusion appears as 8.0% in Table 1 (detection grid), 8.8% in Table 2 (five-model tail-criterion average), and 9.2% here (per-model values).

A.4 Cross-Generator Transfer

Table 5 reports the GPT-2 → modern transfer results of Section 5.5.

Target	ID baseline	GPT-2 only	Δ
Llama-3.1 (8B)	0.991	0.987	+0.004
Mistral-7B	0.956	0.954	+0.002
Qwen 2 (7B)	0.940	0.943	−0.003
Deepseek (7B)	0.956	0.952	+0.004
Falcon 2 (11B)	0.907	0.908	−0.001
Mean	**0.950**	**0.949**	**+0.001**

Table 5: GPT-2 → modern transfer (RF AUC-ROC). A classifier trained only on GPT2-XL outputs is evaluated on each modern open model alongside a target-specific in-distribution baseline.

A.5 Matched Training-Set Comparison

Table 6 reports the matched training-set comparison of Section 5.5.

Held-out M	ID-single	ID-pooled	LOMO	Δ
Deepseek	0.956	0.927	0.929	−0.001
Falcon 2	0.907	0.887	0.893	−0.006
GPT2-XL	0.952	0.925	0.929	−0.005
Llama-3.1	0.991	0.963	0.966	−0.003
Mistral-7B	0.956	0.929	0.935	−0.006
Qwen 2	0.940	0.917	0.922	−0.005
Mean	**0.950**	**0.925**	**0.929**	**−0.004**

Table 6: Matched training-set comparison (RF AUC-ROC). Same test set per held-out model; only training composition varies. Per-classifier statistics in Appendix A.13.

A.6 Case Study: Quality Versus Detectability

Table 7 contrasts continuations of the same prompt, referenced from Section 5.6.

A.7 Extended Related Work

The detection of machine-generated text has attracted significant attention as LLM capabilities have advanced (Jawahar et al., 2020). Existing approaches fall into three main categories. *Zero-shot statistical methods* exploit properties of model probability distributions: DetectGPT (Mitchell et al., 2023) identifies machine text via negative curvature in the log-probability landscape, while likelihood-based approaches exploit differences in token predictability (Gehrmann et al., 2019; Ippolito et al., 2020). *Trained classifiers* distinguish human from machine text using labeled corpora (Solaiman et al., 2019; Uchendu et al., 2020; Tang et al., 2024), but often struggle to generalize across models and domains. *Watermarking methods* embed statistical signatures during generation (Kirchenbauer et al., 2023), though these can be undermined by paraphrasing attacks (Krishna et al., 2023). Our analysis connects these statistical detection signals to the tokens removed by likelihood-based truncation.

Decoding strategies strongly shape the properties of generated text (Wiher et al., 2022; Garces Arias et al., 2025c; Frisoni et al., 2026). Prior work has largely focused on downstream task performance (e.g., summarization), whereas we examine human-token exclusion and detectability across models and domains. Holtzman et al. (2020) show that maximization-based decoding leads to degeneration, motivating nucleus (top-p) sampling to truncate the unreliable tail. Earlier work introduced top-k sampling (Fan et al., 2018), while contrastive search balances likelihood with diversity (Su and Collier, 2023). From an information-theoretic perspective, locally typical sampling targets tokens with information content close to conditional entropy (Meister et al., 2023), drawing on accounts of information density and communicative efficiency (Jaeger, 2010; Gibson et al., 2019). We extend this line of work by linking truncation directly to detectability: mechanisms that improve fluency by excluding low-probability tokens also introduce systematic deviations from human text.

Several studies bear directly on our analysis. Ippolito et al. (2020) show that decoding choices tuned to fool human readers introduce statistical regularities that make automatic detection easier; we account for that trade-off mechanistically, tracing it to the tokens truncation removes. Hewitt et al. (2022) cast truncation as desmoothing and propose η- and ϵ-sampling to recover the support of the true distribution more judiciously; we take the complementary, diagnostic view, asking what such truncation excludes relative to human text, and evaluate their samplers directly against our exclusion benchmark (Section 5.1). Finlayson et al. (2024) use the softmax bottleneck to derive a truncation criterion and evaluate it in pilot generation experiments. Our diagnostic measures human-token exclusion and its relationship with detectability. Closest to our setting, Dubois et al. (2025) show that small changes in sampling parameters can sharply alter detector accuracy; our results concur that decoding governs detectability, and add a mechanistic account in terms of token exclusion, a linguistic decomposition of the excluded set, and coverage of proprietary and non-Transformer models.

Related approaches also address the predictive distribution and its evaluation. Variance-Calibrated Modulation reshapes logits before truncation using contextual relevance and a variance-scaled repetition penalty (Ding et al., 2026a). Credal-set analysis measures gaps between model outputs and human creative variation (Garces Arias et al., 2025d), while multicriteria evaluation compares decoding strategies across quality dimensions (Garces Arias et al., 2025b). These perspectives complement exclusion rates with distributional and quality-based assessments.

Psycholinguistics and information theory offer complementary accounts of language use. Surprisal theory relates processing difficulty to negative log-probability (Hale, 2001; Levy, 2008), with evidence across eleven languages (Wilcox et al., 2023). Speech-production models distinguish conceptual preparation, lexical selection, and articulation (Levelt, 1989; Indefrey and Levelt,

Source	Continuation (excerpt)	P(Machine)
Llama-3.1, beam (3)	"he served as Commanding Officer of the 1st Battalion, The Queen's Royal Regiment [...] Commanding Officer of the 2nd Battalion, The Queen's Royal Regiment [...] of the 3rd Battalion [...]"	1.00
Qwen 2, CS (1.0, 20)	"Headelm oversa Operation Agulhene535af69, to reoccupyo portug 1 st colonial outst a nces in AFRica coun trie s to preva l sepa rabli st mov [...]"	0.88

Table 7: Continuations of the same WikiText prompt (Headlam's service record), with the Random Forest estimate of machine authorship. Degenerate (repetitive) text is highlighted in purple; erratic (incoherent) text in brown. Both machine continuations receive high estimated machine probabilities (1.00 and 0.88), despite exhibiting different forms of poor quality. These individual scores are not measures of population-level detection performance or calibrated certainty. Full case study in Table 24 (Appendix A.21).

2004). Information-density research examines how speakers distribute information across an utterance (Jaeger, 2010), while Shannon (1951) estimates the entropy and redundancy of English. These perspectives motivate studying the information retained by decoding rules.

A.8 Token Category Taxonomy

Table 8 formalizes the four-category taxonomy introduced in Section 2.

	Human Includes	Human Excludes
Model Includes	**Shared Inclusion** $\mathcal{T}_{\mathcal{H}}(\mathbf{x}_{<t}) \cap \mathcal{T}_{\theta}(\mathbf{x}_{<t})$ Tokens that are both statistically probable for the model and communicatively appropriate for humans	**Model Overreach** $\mathcal{E}_{\mathcal{H}}(\mathbf{x}_{<t}) \cap \mathcal{T}_{\theta}(\mathbf{x}_{<t})$ Tokens included by the model's strategy but excluded by humans as communicatively inappropriate
Model Excludes	**Blind Spot** $\mathcal{T}_{\mathcal{H}}(\mathbf{x}_{<t}) \cap \mathcal{E}_{\theta}(\mathbf{x}_{<t})$ *(Focus of this paper)* Tokens that are communicatively valuable, yet are excluded by the likelihood-based truncation	**Shared Exclusion** $\mathcal{E}_{\mathcal{H}}(\mathbf{x}_{<t}) \cap \mathcal{E}_{\theta}(\mathbf{x}_{<t})$ Tokens rejected by both systems (e.g., rarely observed in training data, typically due to grammatical or semantic unsuitability)

Table 8: Taxonomy of token categories based on model and human selection mechanisms. The *blind spot* contains tokens that humans may select for communicative purposes, but that likelihood-based truncation systematically excludes due to low probability.

A.9 Extended Discussion

A.9.1 Summary of Evidence

Our results provide evidence consistent with the hypothesis that the detectability of machine-generated text arises from a distributional mismatch introduced at decoding time, when likelihood-based truncation removes low-rank tokens that human writers use and that the model itself assigns nonzero probability. Under typical truncation settings, 8–18% of human-selected tokens would be excluded, demonstrating that the blind spot is empirically substantial (**RQ1**). Machine text exhibits systematically higher predictability than human text, and this predictability serves as the primary signal enabling detection with simple classifiers (**RQ2**). Stricter truncation increases both the exclusion rate and detectability; relaxing truncation reduces both. Decoding parameters affect detectability, while scale and architecture also show effects in some analyses (**RQ3**). These patterns are consistent with the mechanistic explanation that likelihood-based truncation creates a systematic blind spot. Finally, low detectability provides limited information about perceived quality (**RQ4**).

A.9.2 Why Detection Generalizes Across Models

Cross-model transfer shows that the two features capture regularities shared by multiple generators. Within-family truncation-set overlap and the GPT-2 POS asymmetry provide complementary evidence about shared candidates and the linguistic composition of excluded tokens. Distinguishing their contributions to transfer is a direction for future work.

A.9.3 Limitations of Current Approaches

The weak association between detectability and human quality ratings highlights the need for separate quality evaluation. Current decoding strategies face a trade-off: strict truncation produces coherent but detectable text (restricted to high-probability tokens), while permissive truncation produces less detectable but often incoherent text (including inappropriate low-probability tokens alongside appropriate ones). The core problem is that likelihood-based truncation cannot distinguish between two types of low-probability tokens: those that are con-

textually appropriate (and would be selected by a human pursuing specific communicative goals) and those that are contextually inappropriate. Permissive truncation admits both; strict truncation excludes both. Addressing this limitation would require mechanisms that can identify contextually appropriate tokens regardless of their probability ranking. Probability-floor criteria are a partial step: they widen the pool where the distribution is flat and recover much of the excluded tail (Section 5.1), but they still select by probability alone and accordingly narrow rather than close the blind spot.

A.10 Practical Implications

For detection research. Predictability and diversity suffice when the decoding configuration is known and degrade gracefully across generators and even across unseen generating models (Section 5.5). Classifier confidence, however, does not track perceived quality (Section 5.6); detection systems built on similar features should not equate low classifier confidence with improved generation quality, and should report calibration separately for coherent and incoherent regimes.

For decoding research. The limitation is not truncation itself but the truncation criterion: likelihood alone cannot separate rare but apt tokens from rare and unsuitable ones. The exclusion rate, together with its part-of-speech decomposition (Section 5.3), provides a quantitative benchmark for evaluating tail-aware sampling methods; probability-floor criteria already dominate nucleus sampling on this benchmark (Section 5.1).

For alignment. The evaluated proprietary systems remain detectable, but their training procedures and decoding settings are not independently controlled. These results motivate controlled comparisons of aligned and unaligned checkpoints under identical decoding conditions.

A.11 Human Token Exclusion Rates

Tables 9 and 10 quantify the magnitude of the blind spot discussed in Section 5.1. Exclusion is computed entirely within each model's own tokenizer: each text tokenized by that model, ranks taken under its distribution, exclusion judged against its truncation set; ranks are never compared across vocabularies. We average only the resulting per-model rates, which are unitless.

Truncation Setting	**Exclusion Rate (%)** Book	News	Wiki	**Avg. (95% CI)**
Top-k Sampling				
$k = 1$	49.5	47.8	51.3	49.5 [47.5, 51.9]
$k = 5$	25.0	22.5	27.4	25.0 [23.1, 27.4]
$k = 10$	18.1	15.7	20.4	18.1 [16.3, 20.2]
$k = 20$	12.9	10.8	15.2	13.0 [11.4, 14.8]
$k = 50$	8.0	6.5	9.9	8.1 [6.9, 9.5]
$k = 100$	5.4	4.3	7.0	5.6 [4.7, 6.7]
Top-p (Nucleus) Sampling				
$p = 0.6$	28.1	28.5	29.6	28.7 [27.7, 29.8]
$p = 0.8$	15.0	16.1	15.8	15.6 [14.9, 16.5]
$p = 0.9$	7.4	8.7	7.9	8.0 [7.4, 8.8]
$p = 0.95$	3.5	4.5	3.7	3.9 [3.4, 4.6]
$p = 0.99$	0.6	0.9	0.8	0.8 [0.6, 1.0]

Table 9: Human token exclusion rates $\mathcal{R}_{\text{excl}}$ across truncation configurations. Values are averaged over five language models (GPT2-XL, Qwen2-7B, Mistral-7B-v0.3, Llama-3-8B, Falcon2-11B) and all texts in each dataset split; 95% confidence intervals are computed over this model $\times$ text distribution. Rates are computed within each model's own tokenizer; ranks are never compared across vocabularies, and only the resulting unitless rates are averaged.

Rank Range	**Book**	**News**	**Wiki**
$[1, 5]$	75.0	77.5	72.6
$[6, 10]$	6.9	6.8	7.0
$[11, 20]$	5.2	4.9	5.3
$[21, 50]$	4.9	4.3	5.2
$[51, 100]$	2.6	2.2	2.9
> 100	5.4	4.3	7.0
Median Rank	1.6	1.2	1.8
Mean Rank	75.5	50.5	155.0

Table 10: Distribution of human token ranks according to language model probability distributions, averaged across five models (GPT2-XL, Qwen2-7B, Mistral-7B-v0.3, Llama-3-8B, Falcon2-11B) and all texts in each dataset.

A.12 Detection Performance by Classifier and Model

Table 11 reports aggregate classification performance across the full experimental grid, referenced from Section 5.4; Tables 12 and 13 break performance down by generating model.

Classifier	**AUC-ROC** Mean	Std	**F1-Score** Mean	Std
RF	0.974	0.054	0.953	0.071
LR	0.968	0.068	0.949	0.077
NB	0.972	0.054	0.949	0.069

Table 11: Classification performance across all configurations (n = 1,062 experiments), averaged over all models, datasets, and decoding configurations.

Model	RF AUC	LR AUC	NB AUC
Llama-3.1	0.993 ± 0.007	0.994 ± 0.006	0.993 ± 0.007
GPT2-XL	0.980 ± 0.022	0.979 ± 0.024	0.978 ± 0.026
Mistral-7B	0.980 ± 0.025	0.981 ± 0.024	0.979 ± 0.028
Deepseek	0.976 ± 0.042	0.952 ± 0.069	0.968 ± 0.047
Qwen 2	0.969 ± 0.035	0.964 ± 0.039	0.967 ± 0.033
Falcon 2	0.945 ± 0.041	0.938 ± 0.050	0.948 ± 0.044

Table 12: Average detectability by model (ordered by RF AUC-ROC). GPT2-XL is more detectable than some recent models, while Llama-3.1 has the highest mean AUC in this grid.

Model	Parameters	Avg. AUC-ROC
GPT2-XL	1.5B	0.979
Falcon 2	11B	0.944
Qwen 2	7B	0.967
Mistral-7B	7B	0.980
Deepseek	7B	0.965
Llama-3.1	8B	0.993

Table 13: Model size versus detectability. No clear correlation emerges between parameter count and detection rates.

A.13 Cross-Model Generalization: Statistical Details

This appendix reports per-classifier statistics for the matched training-set comparison and the GPT-2 $\rightarrow$ modern transfer experiments described in Section 5.5.

Setup details. For both experiments, classifiers are trained on data pooled across all decoding configurations (rather than per-cell as in Section 4.5), to enable cross-model comparison on the same test set. The 80/20 train/test split is performed at the prompt level (rather than the row level) to prevent leakage. RF uses 500 trees with `class_weight=balanced`; LR uses `max_iter=2000` with `class_weight=balanced`; NB is Gaussian. All experiments use the same random seed and the same global train/test prompt partition.

Matched comparison: per-classifier statistics. The quantity reported is $\Delta = \text{AUC}_{\text{ID-pooled}} - \text{AUC}_{\text{LOMO}}$, averaged across the six held-out models. 95% confidence intervals are computed from the per-model Δ values using the t-distribution.

Classifier	Mean Δ	95% CI	Sign test	Paired $t(5)$	p
RF	-0.0044	$[-0.0063, -0.0025]$	$p = 0.031$ (6/6 negative)	$t = -5.90$	0.002
NB	-0.00002	$[-0.0002, +0.0002]$	$p = 0.69$ (2/6 negative)	$t = -0.21$	0.84
LR	-0.0024	$[-0.0195, +0.0147]$	$p = 1.00$ (3/6 negative)	$t = -0.36$	0.73

Table 14: Per-classifier statistics for the matched cross-model comparison. Across all three classifiers, the 95% CI for Δ excludes any positive value larger than 0.015; for RF the entire interval is negative, while the NB interval spans zero.

GPT-2 $\rightarrow$ modern transfer: per-classifier statistics. The quantity reported is $\Delta = \text{AUC}_{\text{ID-target}} - \text{AUC}_{\text{GPT-2}\rightarrow\text{target}}$, averaged across the five modern open targets.

Classifier	Mean Δ	95% CI	Paired $t(4)$	p
RF	$+0.0015$	$[-0.0022, +0.0051]$	$t = +1.12$	0.33
NB	$+0.0005$	$[-0.0013, +0.0023]$	$t = +0.75$	0.49
LR	$+0.0090$	$[-0.0233, +0.0412]$	$t = +0.77$	0.48

Table 15: Per-classifier statistics for the GPT-2 $\rightarrow$ modern transfer. The difference between target-specific in-distribution training and GPT-2-only training is not statistically detectable for any classifier.

Interpretation. Two interpretive points are worth emphasizing. First, absolute AUC values in the pooled-training setting are lower than the per-cell AUCs reported in Section 5.4 (mean RF ID baseline 0.950 vs. per-cell mean 0.974). This is expected: pooling across decoding configurations introduces within-class variance from the full hyperparameter sweep, which a per-cell classifier never sees. The relevant quantity for assessing cross-model generalization is the *difference* between matched conditions on the same test set, not the absolute level. Second, the LR confidence interval is wider than RF and NB because LR's linear decision boundary is suboptimal for the pooled feature distribution; this manifests in higher per-cell variance in Δ, not in a systematic deviation from zero.

A.14 Class Balance Robustness

Because the corpus contains $\sim$1.8M machine-generated texts and 5,261 human references, we assessed sensitivity to class imbalance. We repeated the matched cross-model comparison with machine examples downsampled to the human count (5,261 each) and observed essentially unchanged AUC-ROC: mean changes were ≤ 0.001 for NB and RF, and ≤ 0.018 for LR (the slightly larger LR shift is itself within the within-classifier variance reported in Table 14). This is consistent with AUC-ROC's

invariance to class prior and indicates that the imbalance does not drive the reported separability.

A.15 Domain-Specific Patterns

Dataset	RF AUC	LR AUC	NB AUC
BookCorpus	0.982 ± 0.030	0.976 ± 0.039	0.979 ± 0.034
WikiNews	0.960 ± 0.043	0.951 ± 0.055	0.960 ± 0.047
WikiText	0.980 ± 0.023	0.976 ± 0.028	0.977 ± 0.027

Table 16: Detection performance by dataset domain. WikiNews shows the lowest average detectability.

Dataset	Predictability		Diversity	
	Mean	Var	Mean	Var
BookCorpus	-2.94	0.104	94.98	25.54
WikiNews	-2.41	0.077	93.84	18.18
WikiText	-2.83	0.155	93.50	34.49

Table 17: Human text statistics across domains.

WikiNews shows both the lowest detectability and the highest human predictability (-2.41), suggesting that formulaic, factual writing is inherently more predictable and thus easier for likelihood-based generation to approximate. BookCorpus shows the highest detectability, consistent with creative writing requiring more tokens from outside the high-probability region. The inverse relationship between human predictability and detectability across domains supports the framework: domains where human text already favors high-probability tokens leave less room for the distributional mismatch that enables detection.

A.16 Extreme Configurations

Tables 18 and 19 list the most and least detectable configurations in the grid.

Model	Strategy	Config.	Domain	Classifier	AUC
Deepseek	CS	(1.0, 15)	Book	LR	1.000
Deepseek	CS	(1.0, 15)	Book	RF	1.000
Deepseek	CS	(1.0, 50)	Book	LR	1.000
Deepseek	CS	(1.0, 50)	Book	RF	1.000
Deepseek	beam	15	Book	LR	1.000
Deepseek	beam	15	Book	NB	1.000
Deepseek	beam	15	Book	RF	1.000
Deepseek	beam	20	Book	LR	1.000
Deepseek	beam	20	Book	RF	1.000

Table 18: Selected evaluations achieving the highest recorded AUC in the grid; contrastive-search examples use $k \geq 3$. Many evaluations tie at the stored precision; rows are examples from that tied set, not a unique ranking.

Model	Strategy	Config.	Domain	Classifier	AUC
Deepseek	CS	(0.6, 20)	Book	LR	0.529
Falcon 2	temp	1	News	LR	0.532
Qwen 2	CS	(0.8, 3)	Wiki	LR	0.537
Deepseek	CS	(0.6, 15)	Wiki	LR	0.569
Qwen 2	CS	(0.8, 3)	Wiki	NB	0.570
Falcon 2	CS	(0.8, 3)	News	LR	0.608
Deepseek	CS	(0.6, 3)	News	LR	0.610
GPT2-XL	temp	1	News	LR	0.615
Deepseek	CS	(0.8, 5)	Book	LR	0.616
GPT2-XL	temp	1	News	RF	0.617

Table 19: Ten lowest-AUC classifier/configuration evaluations in the 59-configuration grid. A configuration can occur for more than one classifier.

Low-AUC configurations occur for several models and include contrastive search and temperature sampling. These examples show the importance of reporting the decoding setting and classifier alongside the model.

A.17 Proprietary Models

Figure 4 and Table 20 present detection results for GPT-3.5-turbo and Claude-3-Haiku, referenced from Section 5.5.

The logistic regression coefficients quantify feature contributions:

$$\log\left(\frac{P(\text{Human})}{P(\text{Machine})}\right) = 6.24 + 23.19 \cdot \text{Div} - 12.88 \cdot \text{Pred}. \quad (4)$$

Both coefficients are significant ($p < 2 \times 10^{-16}$). Higher lexical diversity is associated with human classification; higher predictability is associated with machine classification, consistent with the prediction that likelihood-based truncation concentrates token selection in high-probability regions.

GPT-3.5-turbo and Claude-3-Haiku remain highly detectable in this evaluation (AUC-ROC > 0.98). Because the study does not compare aligned and unaligned versions under identical decoding conditions, these observations do not identify the effect of alignment on detectability.

A.18 Full Hyperparameter Results

Table 21 presents detection performance and text properties for 48 non-temperature configurations: beam search, top-k, top-p, and contrastive search with $k \geq 3$.

A.19 Effect of Architecture and Scale on AUC-ROC

To isolate the effects of model scale and architecture on detectability, we conducted a hierarchical

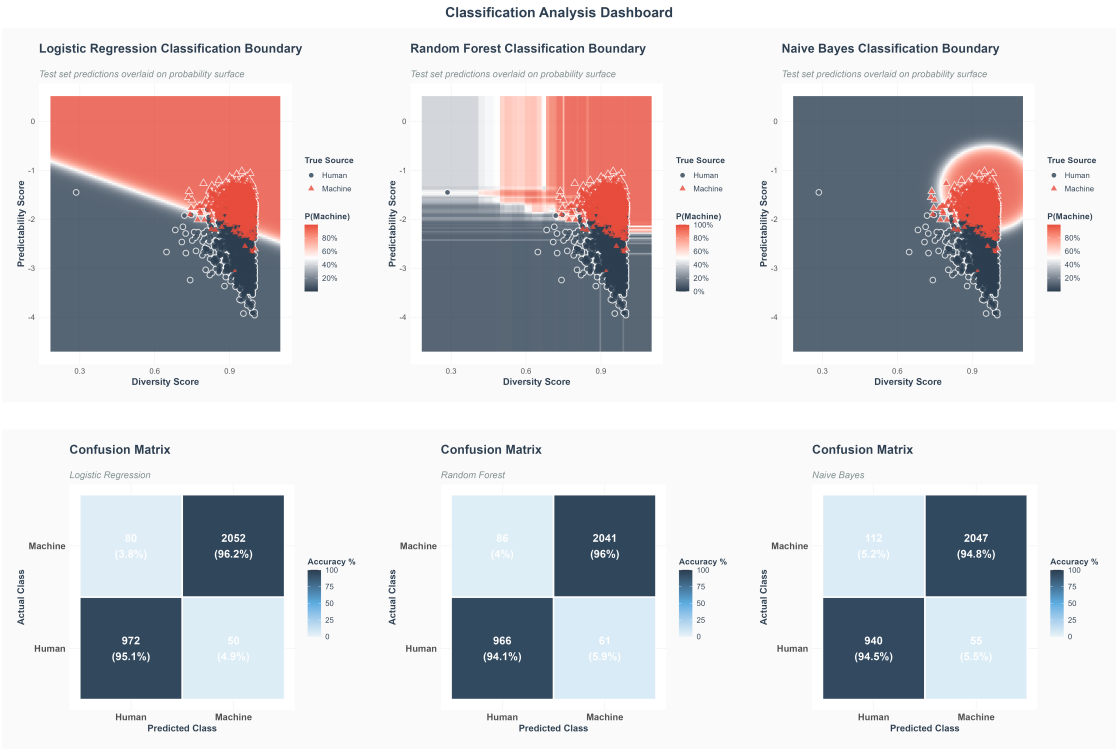


Figure 4: Binary classification performance for proprietary models GPT-3.5-turbo and Claude-3-Haiku. Despite advanced training procedures and alignment, both remain highly detectable using only predictability and diversity features.

Classifier	Performance Metrics						
	Acc.	Prec.	Rec.	F1	Spec.	AUC-ROC	AUC-PR
RF	0.953	0.960	0.971	0.965	0.918	0.983	0.988
LR	**0.959**	**0.963**	**0.976**	**0.969**	**0.924**	**0.989**	**0.994**
NB	0.947	0.948	0.974	0.961	0.894	0.988	0.993

Table 20: Detection performance for proprietary models (GPT-3.5-turbo and Claude-3-Haiku combined).

regression analysis controlling for confounding factors; Figure 5 visualizes the resulting effects.

Scale Analysis. We regressed AUC-ROC on log-transformed parameter count while conditioning on model family, dataset, and decoding strategy. Table 22 reports the results. The coefficient of $\log(\text{params})$ is not statistically significant in family-specific regressions for three of four model families: Qwen2 ($\beta = -0.020$, $p = 0.229$), Mamba ($\beta = -0.013$, $p = 0.108$), and RWKV ($\beta = 0.000$, $p = 0.468$). Only GPT2 shows a significant within-family effect ($\beta = -0.052$, $p = 0.015$). Similar patterns hold for Logistic Regression and Naive Bayes classifiers. These results indicate that *within model families*, increasing scale does not reliably reduce detectability.

Architecture Analysis. Controlling for scale, dataset, and decoding strategy, we find significant differences across architectures. Non-Transformer architectures (Mamba, RWKV) exhibit significantly higher detectability than Transformer baselines ($\Delta\text{AUC} = 0.180$, $p < 0.001$). This suggests that high detectability is *not unique to Transformer architectures*; rather, it may persist in non-Transformer models when generation relies on likelihood-based truncation.

Strategy	Config.	Classification Metrics				Text Properties	
		Prec.	Rec.	F1	AUC-ROC	Pred.	Div.
Beam Search	3	0.987	0.993	0.989	0.996	-1.67 ± 1.15	59.20 ± 40.76
	5	0.986	0.993	0.990	0.995	-1.68 ± 1.14	58.72 ± 41.08
	10	0.987	0.994	0.991	0.996	-1.67 ± 1.14	57.57 ± 41.74
	15	0.987	0.992	0.990	0.996	-1.67 ± 1.15	56.91 ± 42.14
	20	0.989	0.994	0.992	0.997	-1.69 ± 1.13	56.40 ± 42.43
	50	0.989	0.995	0.992	0.997	-1.66 ± 1.15	54.91 ± 43.41
Contrastive Search	(0.2, 3)	0.968	0.985	0.976	0.989	-1.89 ± 1.00	64.18 ± 37.60
	(0.2, 5)	0.973	0.985	0.979	0.991	-1.82 ± 1.06	65.38 ± 36.84
	(0.2, 10)	0.970	0.982	0.976	0.990	-1.87 ± 1.01	66.72 ± 36.06
	(0.2, 15)	0.966	0.981	0.973	0.988	-1.87 ± 0.98	67.51 ± 35.69
	(0.2, 20)	0.972	0.987	0.979	0.990	-1.84 ± 1.03	68.35 ± 35.34
	(0.2, 50)	0.963	0.980	0.971	0.985	-1.87 ± 1.01	70.97 ± 34.03
	(0.4, 3)	0.960	0.972	0.966	0.985	-2.17 ± 0.88	73.71 ± 30.71
	(0.4, 5)	0.941	0.963	0.951	0.973	-2.10 ± 0.94	76.79 ± 28.30
	(0.4, 10)	0.930	0.956	0.943	0.966	-2.22 ± 0.89	80.06 ± 25.88
	(0.4, 15)	0.941	0.967	0.954	0.974	-1.99 ± 0.90	82.33 ± 24.17
	(0.4, 20)	0.954	0.965	0.959	0.976	-2.15 ± 0.93	83.82 ± 22.83
	(0.4, 50)	0.948	0.963	0.955	0.969	-2.06 ± 0.87	87.11 ± 19.29
	(0.6, 3)	**0.852**$\star$	0.891	**0.870**$\star$	**0.901**$\star$	-2.27 ± 0.78	81.67 ± 23.77
	(0.6, 5)	0.923	0.943	0.933	0.961	-2.26 ± 0.80	85.05 ± 20.83
	(0.6, 10)	0.886	0.906	0.895	0.934	-2.38 ± 0.76	87.84 ± 18.19
	(0.6, 15)	0.898	0.911	0.904	0.934	-2.25 ± 0.81	88.88 ± 17.35
	(0.6, 20)	0.920	0.938	0.929	0.949	-2.33 ± 0.83	89.42 ± 16.97
	(0.6, 50)	0.934	0.933	0.933	0.962	-2.45 ± 0.85	90.85 ± 16.00
	(0.8, 3)	0.883	0.906	0.893	0.916	-2.90 ± 1.01	87.39 ± 18.49
	(0.8, 5)	0.890	0.912	0.900	0.931	-2.87 ± 1.21	89.31 ± 16.91
	(0.8, 10)	0.941	0.955	0.948	0.966	-3.08 ± 1.04	90.85 ± 15.87
	(0.8, 15)	0.963	0.968	0.965	0.981	-2.98 ± 1.14	91.40 ± 15.49
	(0.8, 20)	0.964	0.968	0.966	0.985	-2.86 ± 1.11	91.72 ± 15.26
	(0.8, 50)	0.988	0.977	0.982	0.994	$-2.86 \pm 0.93\star$	92.04 ± 15.03
	(1.0, 3)	0.973	0.976	0.975	0.990	-3.41 ± 1.30	90.20 ± 16.52
	(1.0, 5)	0.975	0.975	0.975	0.990	-3.43 ± 1.56	91.01 ± 16.00
	(1.0, 10)	0.991	0.989	0.990	0.998	-3.69 ± 1.47	91.70 ± 15.49
	(1.0, 15)	0.997	0.996	0.996	1.000	-3.46 ± 1.44	91.90 ± 15.23
	(1.0, 20)	0.996	0.989	0.993	0.997	-3.37 ± 1.40	91.92 ± 15.23
	(1.0, 50)	0.993	0.990	0.991	0.996	-3.07 ± 1.14	92.12 ± 15.00
Top-k	1	0.961	0.981	0.971	0.988	-2.16 ± 0.77	57.82 ± 41.59
	3	0.930	0.940	0.934	0.967	-2.27 ± 0.63	75.78 ± 27.15
	5	0.919	0.932	0.925	0.957	-2.10 ± 0.82	82.40 ± 20.66
	10	0.885	0.890	0.887	0.929	-2.26 ± 0.67	87.06 ± 15.68
	15	0.876	**0.880**$\star$	0.877	0.924	-2.29 ± 0.62	87.18 ± 16.37
	20	0.898	0.886	0.891	0.929	-2.21 ± 0.71	87.82 ± 15.78
	50	0.924	0.925	0.924	0.955	-2.13 ± 0.74	88.29 ± 16.12
Top-p	0.6	0.940	0.953	0.946	0.978	-2.21 ± 0.66	75.29 ± 28.39
	0.7	0.916	0.934	0.925	0.960	-2.26 ± 0.62	81.22 ± 22.58
	0.8	0.923	0.928	0.926	0.961	-2.22 ± 0.67	86.22 ± 16.64
	0.9	0.914	0.910	0.912	0.949	-2.23 ± 0.66	89.97 ± 11.53
	0.95	0.891	0.890	0.890	0.935	-2.28 ± 0.60	91.39 ± 9.43
Human	–	–	–	–	–	-2.73 ± 0.11	94.11 ± 1.60

Table 21: Detection performance and text properties for 48 selected non-temperature configurations, with $k \geq 3$ for contrastive search. Predictability and diversity summarize the combined human and machine samples within each cell, not machine generations alone; they therefore should not be interpreted as generation-only statistics. Values highlighted in blue indicate lowest detection rates within each strategy. Values highlighted in green indicate configurations closest to human baseline values. Best overall configurations marked with $\star$. Human baseline shown for comparison.

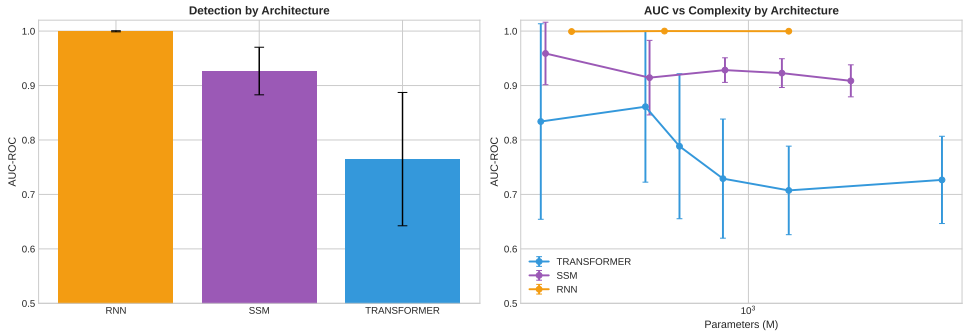


Figure 5: Effect of architecture and scale on AUC-ROC, comparing Transformer- and Non-Transformer-based architectures at different sizes.

Predictor	β	SE	p-value
Scale Effect (within-family)			
log(params)	-0.021	0.009	$0.017^{\dagger}$
Architecture Effect (controlling for scale)			
Non-Transformer vs. Transformer	$+0.180$	0.018	< 0.001

Table 22: Regression coefficients for scale and architecture effects on AUC-ROC (Random Forest classifier). †Pooled estimate; family-specific regressions show non-significant effects for 3 of 4 families (Qwen2: $p = 0.229$; Mamba: $p = 0.108$; RWKV: $p = 0.468$).

Interpretation. Detectability persists across the evaluated model families, but both the pooled scale coefficient and the architecture coefficient are statistically significant. Decoding and model characteristics both contribute to the observed variation in detectability.

A.20 Robustness to the Choice of Reference Model

Our analysis relies on predictability estimates computed using a reference language model (OPT-2.7B). To assess whether our findings depend on this specific choice, we evaluate agreement across multiple judge models spanning different architectures and scales, including OPT-2.7B, Llama-3.1-8B, Qwen2-7B, and GPT2-medium. For each model, we compute predictability scores over the same set of human and machine-generated texts and measure pairwise Spearman rank correlations. As shown in Table 23, we observe consistently high correlations across all model pairs ($\rho \geq 0.88$), indicating strong agreement in the relative ordering of texts by predictability.

	OPT-2.7B	Llama-3.1-8B	Qwen2-7B	GPT2-medium
OPT-2.7B	1.00	0.93	0.91	0.88
Llama-3.1-8B	0.93	1.00	0.94	0.90
Qwen2-7B	0.91	0.94	1.00	0.89
GPT2-medium	0.88	0.90	0.89	1.00

Table 23: Spearman rank correlation (ρ) between predictability scores computed using different judge models.

A.21 Case Study

Table 24 extends the case study of Section 5.6 with longer excerpts on the same prompt.

A.22 Contrastive Search Parameter Interactions

Figure 6 shows AUC-ROC heatmaps for contrastive search, illustrating the interaction between the k and α parameters across all three datasets. Darker regions indicate lower detectability; the least detectable cells concentrate at intermediate α with moderate k, consistent with Table 19.

A.23 AUC-ROC Evolution Across Configurations

Figure 7 traces the evolution of AUC-ROC across hyperparameter values for multiple decoding strategies; the variation across settings shows how decoding choices affect detectability.

A.24 Token Set Overlap at Different Truncation Levels Across Varying Scale

Figure 8 visualizes within-family truncation-set overlap using the Jaccard distance, referenced from Sections 5.4 and 5.5.

Overlap Discussion. We examine whether models within the same architectural family converge on similar truncation sets. Overlap describes how candidate sets change across model sizes; it does

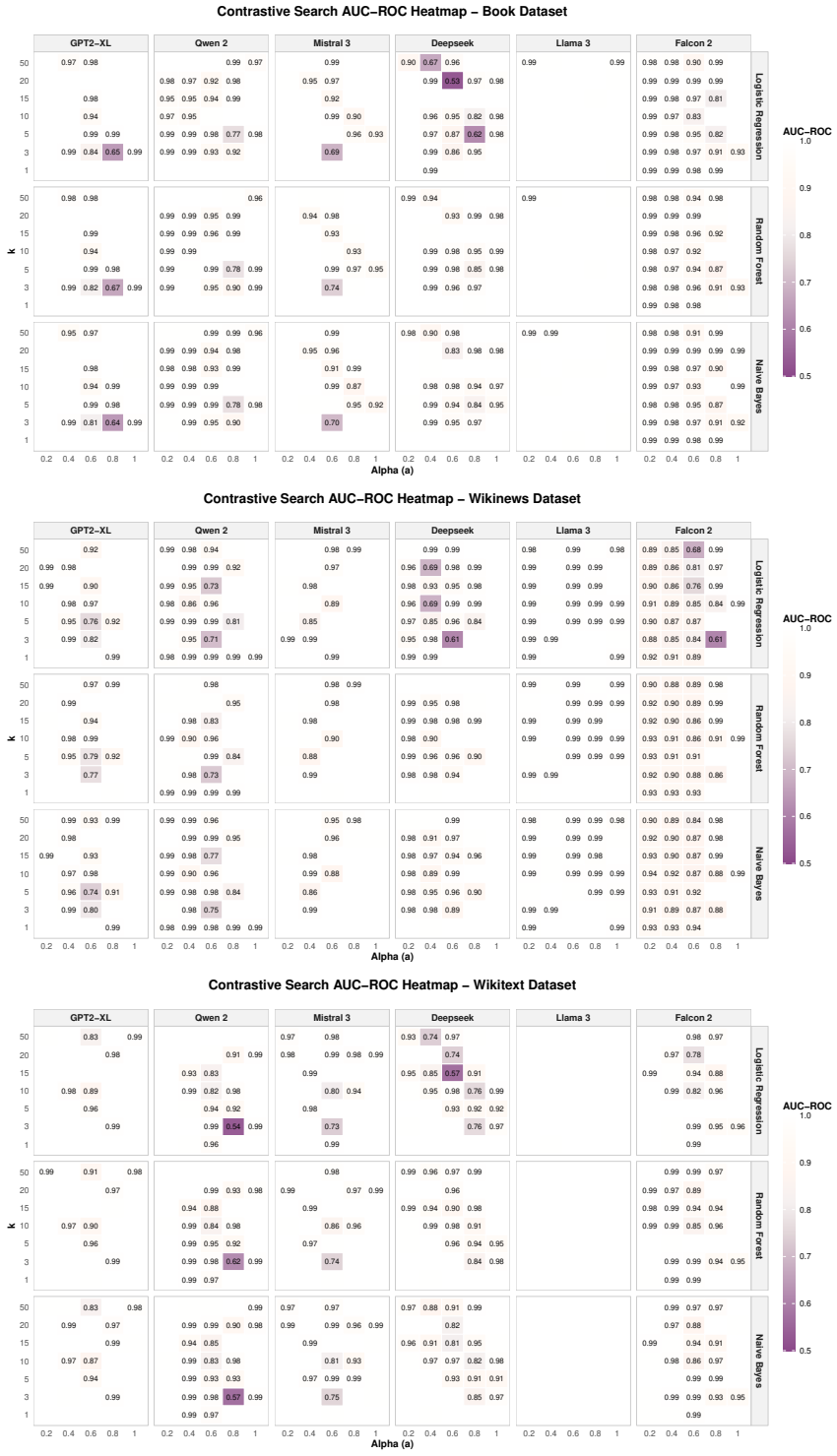


Figure 6: AUC-ROC heatmaps for contrastive search showing interactions between k and α parameters. Darker regions indicate lower detectability. Facet labels "Llama 3" and "Mistral 3" denote the Llama-3.1 and Mistral-7B checkpoints of Section 4.

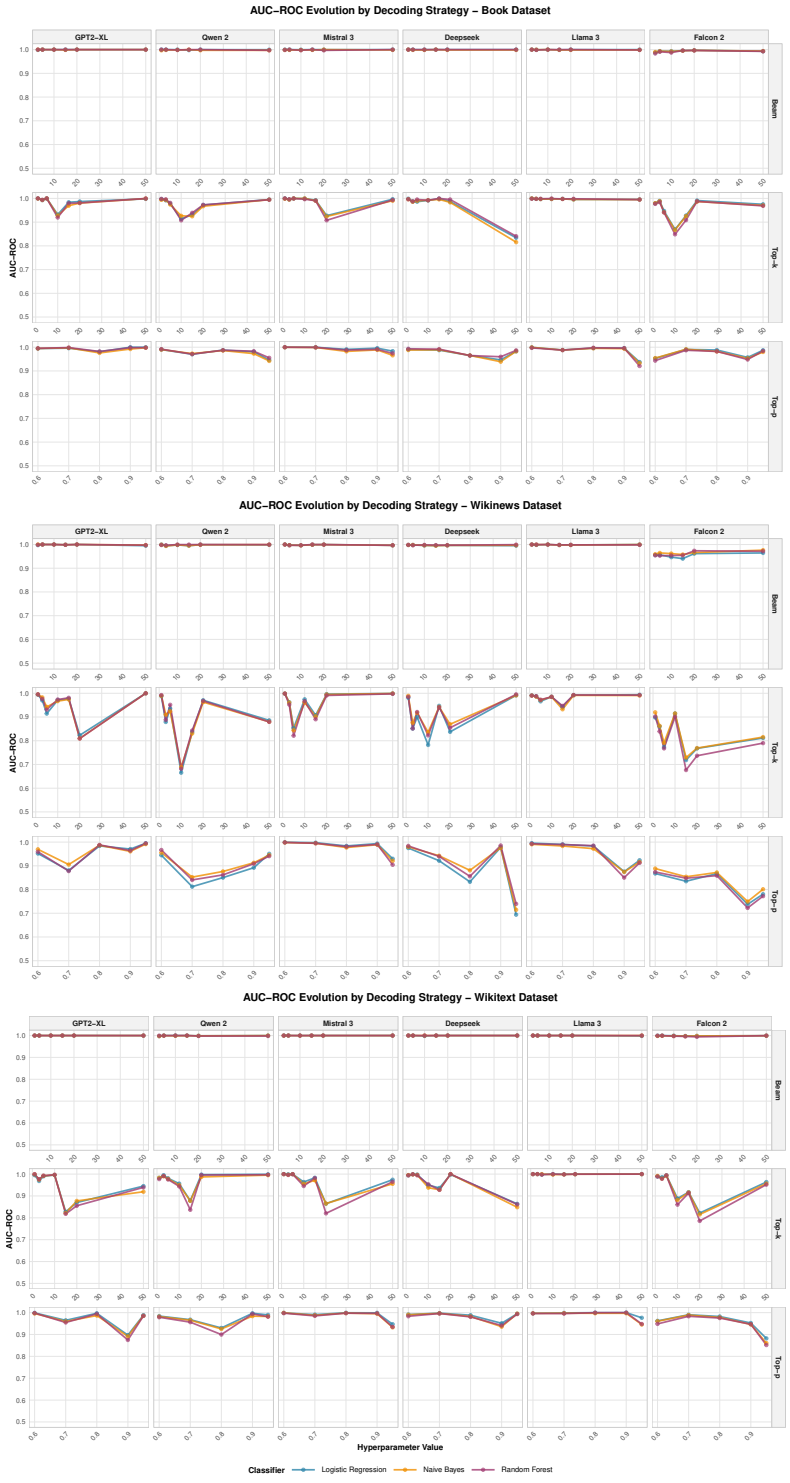


Figure 7: Evolution of AUC-ROC across hyperparameter values for multiple decoding strategies, models, and datasets. Larger k/p values generally reduce detectability, consistent with the hypothesis that relaxing truncation reduces the distributional mismatch. Facet labels "Llama 3" and "Mistral 3" denote the Llama-3.1 and Mistral-7B checkpoints of Section 4.

Decoding Method	Prompt: "Headlam served as Officer Commanding North @-@ Western Area in 1946, and as Director of Training from 1947 to 1950. In 1950 - 51,"	P(Machine)
Qwen 2 CS (1.0, 20)	" Headelm oversa Operation Agulhene535af69, to reoccupyo portug 1 st colonial outst a ncesin AFRicacoun trie sto preva l sepa rabli st mov me s . In th es on he wro 4he couthoea2 rsh ipp ublicatin hitherto re t aine o secret ... (text continues)"	0.88
Llama-3.1 beam (3)	"he served as Commanding Officer of the 1st Battalion, The Queen's Royal Regiment (West Yorkshire Regiment). In 1951 2013 52, he served as Commanding Officer of the 2nd Battalion, The Queen's Royal Regiment (West Yorkshire Regiment). In 1952 2013 53, he served as Commanding Officer of the 3rd Battalion, The Queen's Royal Regiment (West Yorkshire Regiment). In 1953 2013 54 , he served as Commanding Officer of the 4th Battalion, ... (text continues)"	1.00

Table 24: Case study comparing text quality and detectability. The probability of machine authorship ($P(\text{Machine})$) is estimated by Random Forest. Degenerate (repetitive) text is highlighted in purple; erratic (incoherent) text is highlighted in brown. The examples illustrate different forms of poor-quality machine output.

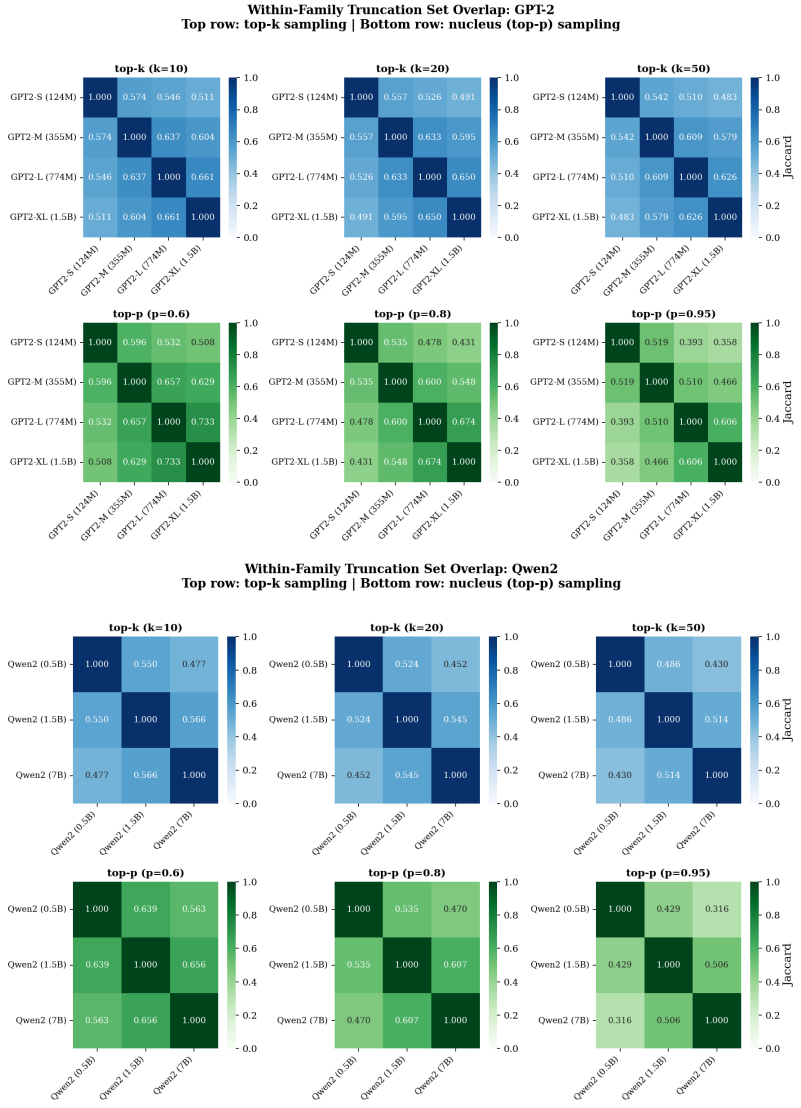


Figure 8: Within-family truncation set overlap, as measured by the Jaccard distance. **Top:** Example for GPT2 (from 124M to 1.5B). **Bottom:** Example for Qwen2 (from 0.5B to 7B) at different truncation schemes: **Blue:** Top-k sampling with $k \in \{10, 20, 50\}$. **Green:** Top-p sampling with $p \in \{0.6, 0.8, 0.95\}$.

not by itself measure recovery of human-selected tokens. Within-family overlap is consistently high across both families. At $k = 10$, GPT-2 models exhibit a mean overlap of 0.589 ($\pm$0.057), while Qwen2 models show 0.531 ($\pm$0.058). Overlap decreases with the size gap between models: for GPT-2, adjacent sizes achieve the highest overlap (0.661 for GPT2-L to GPT2-XL), while the largest gap yields the lowest (0.511 for GPT2-S to GPT2-XL). The same pattern holds for Qwen2, with overlap ranging from 0.566 (adjacent) to 0.477 (largest gap). This gradient indicates that scaling induces gradual distributional drift, but does not fundamentally restructure which tokens receive high probability. Under nucleus sampling, overlap decreases as p increases, reflecting the expanding nucleus size: at $p = 0.95$, mean overlap is 0.475 (GPT-2) and 0.417 (Qwen2), yet remains substantial.

A.25 Detectability–Quality Dissociation

Figure 9 plots human quality ratings against estimated machine probability, quantifying the dissociation reported in Section 5.6.

A.26 Linguistic Composition of the Blind Spot

Sample and checkpoints. We sampled 200 unique prompt/reference pairs without replacement from each of the BookCorpus, WikiNews, and WikiText reference checkpoints in our public repository. The available files contain 1,939, 2,000, and 1,300 unique pairs, respectively; this POS subset is distinct from the full detection corpus. The saved input manifest identifies every selected pair. We use `gpt2`, `gpt2-medium`, and `gpt2-large` (124M, 355M, and 774M parameters), with float32 inference. Their shared tokenizer holds subword segmentation fixed across the three sizes, allowing a within-family analysis. Extending the analysis to other model families remains future work.

Scoring and alignment. We remove an explicit leading `<|endoftext|>` marker from the prompt, concatenate the original prompt and reference, and tag the combined text with spaCy 3.8.7, `en_core_web_sm` 3.8.0. A fast GPT-2 tokenizer supplies character offsets. We score at most 256 continuation subwords under the full preceding prompt and continuation context; the first unconditioned token, if any, is not scored. A subword receives a POS tag only when all its non-whitespace characters belong to one spaCy token. Whitespace-only tokens and subwords spanning multiple words are excluded from POS denominators. Ranks use the conservative tie convention of Section 2.

Alignment audit. Across the three models, 353,808 continuation-token observations were selected and 346,518 were aligned: 4,689 whitespace-only observations and 2,601 multiword observations were omitted. There were no unmapped or prompt-boundary-crossing observations. Multiword omissions represent 0.55–1.01% of selected tokens by domain; the GPT-2 variants share these alignments. No ties straddled the $k = 10$ or $k = 50$ boundaries. As a conservative sensitivity check, assigning every omitted multiword observation to the function group and treating it as excluded reduces the pooled $k = 10$ ratio to 3.07, leaving its direction unchanged.

Groups and aggregation. Content categories are NOUN, VERB, ADJ, ADV, PROPN, and NUM; function categories are DET, ADP, CCONJ, SCONJ, PRON, AUX, and PART. Punctuation, interjections, symbols, and other tags form a separate group. Reported and plotted rates pool excluded and total aligned subword counts across the fixed model set and sampled pairs. Thus, a word split into multiple subwords contributes multiple observations. A supplementary macro-average, weighting observed POS/model/domain rates equally, gives 47.2% versus 12.8% and a ratio of 3.69 at $k = 10$; this differs from the pooled ratio of 3.53.

Uncertainty and consistency. The 95% percentile interval uses 1,000 bootstrap resamples of prompt/reference pairs within each domain (seed 42); each resampled pair carries its observations for all three models and both word groups. It quantifies passage-sampling uncertainty conditional on these models and automatic tags, not uncertainty over model families or original source documents. Pooled exclusion is higher for content than function subwords in every model/domain cell (Table 25).

Category and domain patterns. At $k = 10$, adverbs (54.5%), verbs (52.5%), adjectives (52.2%), and numerals (47.7%) have the highest exclusion rates among the designated content categories. Proper nouns and nouns have rates of 35.4% and 35.2%. Function-category rates range from 6.7% for particles to 26.0% for subordinating conjunctions. Interjections have a higher rate but only 111 model-token observations and are outside the designated content/function groups.

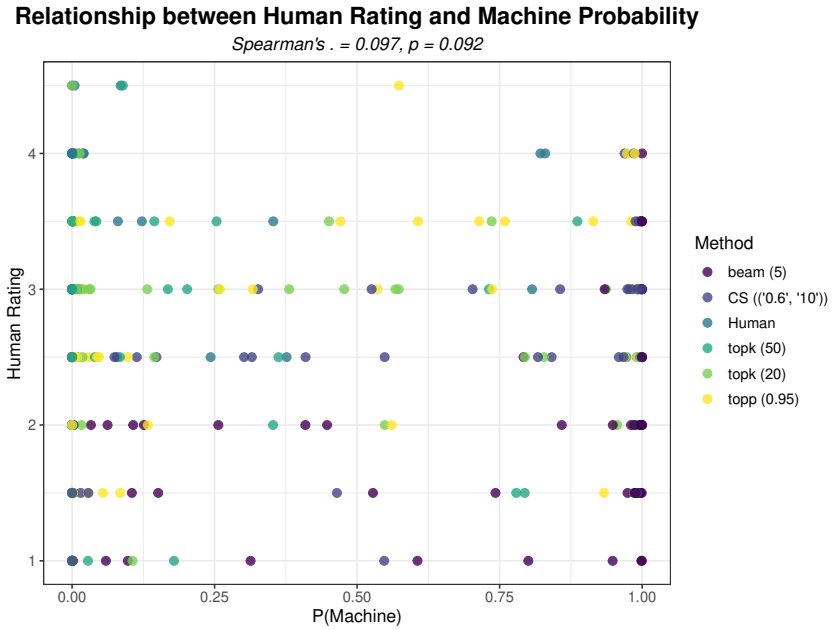


Figure 9: Relationship between human quality ratings (y-axis; Likert scale 1–5) and the estimated probability of machine-generated text (x-axis) across different decoding strategies. The association is weak, with a very small and statistically non-significant Spearman correlation. Human-written texts receive the highest ratings, followed by Top-p sampling, and Top-k sampling. Beam search yields the lowest ratings, with contrastive search performing slightly better. Results were computed from publicly available human evaluation results by Garces Arias et al. (2025a).

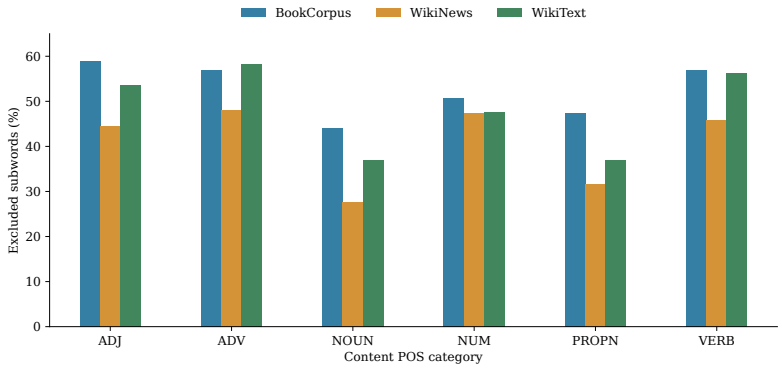


Figure 10: Content-category subword exclusion at $k = 10$ by benchmark domain, pooling counts across the three GPT-2 models. Each domain contributes 200 prompt/reference pairs; the number of aligned subwords varies by domain.

Figure 10 shows domain variation: BookCorpus has the highest pooled exclusion for adjectives, nouns, proper nouns, and verbs, while WikiText is highest for adverbs.

Category prevalence. Across 17 observed POS categories, category prevalence (the share of aligned subword observations) has little linear association with pooled exclusion ($r = -0.094$, $p = 0.719$; Figure 3C). A common category can contain many rare words, so this aggregate comparison does not rule out lexical-frequency effects.

Model	Domain	Content	Function	Ratio
GPT-2 small	Book	55.5	14.7	3.77
	News	40.5	11.7	3.46
	Wiki	48.4	13.0	3.73
GPT-2 medium	Book	50.3	13.3	3.78
	News	34.9	10.7	3.27
	Wiki	42.4	11.6	3.64
GPT-2 large	Book	48.9	12.7	3.84
	News	32.5	9.8	3.33
	Wiki	39.8	10.7	3.73

Table 25: Pooled subword exclusion rates (%) at $k = 10$ within each GPT-2 model/domain cell. Ratios are computed before rounding.