

Formatting and Surface Feature Confounds in AI Text Detection: A Clever Hans Effect

Koel Dutta Chowdhury^{1*} Cristina España-Bonet^{2,3} Josef van Genabith^{3,4}

¹Technical University of Nuremberg, Germany

²Barcelona Supercomputing Center (BSC-CNS), Catalonia, Spain

³DFKI GmbH, Saarland Informatics Campus, Germany

⁴Saarland University, Saarland Informatics Campus, Germany

koel.dutta.chowdhury@utn.de

Abstract

AI-text detectors are typically evaluated on benchmarks that pair human-written text with text generated by language models. However, differences in how these two classes are constructed can introduce surface-form cues that make the benchmark easier to solve. We show that surface form alone strongly separates human from machine-generated text across three detection benchmarks, but the learned surface patterns vary substantially across benchmarks. Detector scores are also substantially predictable from surface form, although detectors retain signal beyond it, and controlled re-formatting reveals partial and detector-specific causal effects. Our findings show that benchmark performance alone does not establish that a detector has identified a stable and transferable signal of machine generation.

1 Introduction

AI-text detectors are systems that distinguish text written by humans from text generated by language models (Ippolito et al., 2020; Wu et al., 2025). They are evaluated on benchmarks that pair human-written text with text generated by one or more language models. When a detector achieves high AUROC on such a benchmark, it is typically assumed to have learned a property specific to machine-generated text (Gehrmann et al., 2019; Su et al., 2023; Mitchell et al., 2023; Hu et al., 2023; Bao et al., 2024; Hans et al., 2024; Xiong et al., 2026).

The human and machine portions of these benchmarks, however, are constructed differently. Human text is collected from existing sources, while machine text is generated with particular models, prompts, and generation settings (Dugan et al., 2024; Wang et al., 2024; Roy et al., 2025). These construction choices can produce systematic differences in surface form, such as paragraph structure, whitespace, punctuation, and length, that predict

the benchmark label independently of any content-level signal. Detectors evaluated on such a benchmark can then achieve high AUROC by relying on surface artifacts of construction, a *Clever Hans* effect (Pfungst, 1911; Geirhos et al., 2020) in which reported accuracy overstates what has actually been detected by the system.

We investigate whether this occurs in current AI-text detection benchmarks. To this end, we introduce a surface-form diagnostic: a linear classifier trained on 86 features that describe how a text is arranged but not what it says. Following the logic of hypothesis-only baselines in natural language inference (Poliak et al., 2018; Gururangan et al., 2018), our diagnostic is designed to show how much of the benchmark label is recoverable from information that should not, by itself, indicate whether a text was generated by a machine. We evaluate it on three benchmarks (RAID, M4, NYT), against ten detectors (eight zero-shot, two supervised), and complement the correlational analysis with cross-benchmark transfer tests for generalizability and a content-preserving paragraph-normalization intervention to isolate a causal effect of surface form on detector scores. Empirically, we find surface form to be strongly predictive of the benchmark label. Our findings are:

① **Surface form separates human and machine text within individual benchmarks.** The strict formatting diagnostic reaches 0.964, 0.956, and 0.995 AUROC on RAID, M4, and NYT, respectively. On RAID, document structure alone reaches 0.932 AUROC, above every detector we evaluate, while a length-only baseline reaches only 0.624.

② **The learned surface patterns are benchmark-specific.** A diagnostic trained on RAID transfers well to held-out RAID domains (0.945 AUROC) but reaches only 0.28 AUROC on M4. The same asymmetry holds across alternative classifiers. Surface form provides a strong signal within a benchmark, but the properties

*Work done while at Saarland University.

associated with machine-generated text differ across benchmarks.

③ **Existing detectors depend on these surface differences to varying degrees.** Surface features explain up to 69% of detector-score variance on RAID. On the subset of AI text with human-like surface form, detector AUROC drops by up to 20%. Merging paragraph breaks in AI samples without changing any words shifts Entropy and LRR scores toward the human distribution but leaves Log-Likelihood and Rank essentially unchanged, providing partial and detector-specific causal evidence.

Taken together, a large fraction of reported detector accuracy on current benchmarks reflects benchmark-specific surface artifacts rather than a stable and transferable signature of machine-generated text.

2 Benchmark Confounds in AI-Text Detection

Recent work suggests that AI-text detectors can rely on signals that are specific to the data on which they are evaluated. Pedrotti et al. (2025) show that shifting machine-generated text toward human writing style substantially reduces detector performance and attribute this behavior to linguistic shortcuts. Borile and Abrate (2025) similarly identify linguistic and domain confounders in supervised detectors that encode dataset-specific rather than task-specific information. More recently, Gao et al. (2026) show that features separating human and machine text can even reverse their effect across evaluation settings. These results suggest that a feature can be highly predictive of machine-generated text in one setting without providing a stable signal in another. This concern is also reflected in recent detector evaluation.

A parallel body of work approaches the question from linguistic analysis. Human and machine-generated text differ in lexical, syntactic, stylistic, and structural properties (Zanotto and Aroyehun, 2025; Guo et al., 2023). However, Attar et al. (2026) find that many linguistic indicators are strongly dependent on the model and domain in which they are measured. We study this problem at the level of benchmark construction, following the logic of diagnostic baselines in NLP (Poliak et al., 2018; Gururangan et al., 2018). Two recent lines of work address these sources of instability directly. Yu et al. (2025) show that detector performance de-

grades as the underlying generators evolve, while Pu et al. (2026) explicitly target generator-specific artifacts that limit generalization to unseen models.

3 Measuring Surface-Form Confounds

We argue that current AI-text detection benchmarks contain *surface-form confounds* (systematic differences in paragraph structure, whitespace, punctuation, and other surface properties between the human and machine portions) that inflate reported detector accuracy without reflecting a stable property of machine-generated text. To support this argument, we measure three quantities: benchmark separability from surface form alone, cross-benchmark transferability of the recovered patterns, and detector sensitivity to surface form.

Benchmark separability. Let a benchmark $\mathcal{B}$ be a labeled set $\{(x_n, y_n)\}_{n=1}^N$, where each text x_n carries a label $y_n \in \{0, 1\}$ indicating whether it is human-written or machine-generated. We define a *surface-form feature map* $\phi : x \mapsto \mathbb{R}^{86}$ that extracts 86 features describing how a text is arranged on the page (paragraph structure, whitespace, punctuation, sentence length, and related properties), computed without any language-model probabilities (see Appendix C). We train a linear classifier

$$g_{\mathcal{B}} : \phi(x) \mapsto \Pr(y = 1 \mid \phi(x)) \quad (1)$$

on $\mathcal{B}$ via 5-fold stratified cross-validation and measure separability by out-of-fold AUROC($g_{\mathcal{B}}$). Following the logic of hypothesis-only baselines in natural language inference (Poliak et al., 2018; Gururangan et al., 2018), $g_{\mathcal{B}}$ is a diagnostic, not a proposed detector: its purpose is to quantify how much of y is recoverable from information that should not, on its own, indicate whether a text is machine-generated. A high AUROC($g_{\mathcal{B}}$) therefore quantifies the surface-form gap between the human and machine portions of $\mathcal{B}$. We additionally consider a *strict formatting* feature map $\phi^{\text{strict}} \subset \phi$ that removes lexical, stylistic, and syntactic features from ϕ , retaining only structural, length, punctuation, and character-ratio properties.

Cross-benchmark transfer. If the surface-form patterns learned by $g_{\mathcal{B}_i}$ reflected a stable property of machine-generated text, $g_{\mathcal{B}_i}$ trained on benchmark $\mathcal{B}_i$ should also separate human from machine text on any other benchmark $\mathcal{B}_j$. To measure this, we train $g_{\mathcal{B}_i}$ on $\mathcal{B}_i$ and evaluate its AUROC on $\mathcal{B}_j$,

and compare within- and across-benchmark transfer:

$$\tau^{\text{within}}(\mathcal{B}_i) = \text{AUROC}(g_{\mathcal{B}_i} \mid \mathcal{B}_i^{\text{held}}), \quad (2)$$

$$\tau^{\text{cross}}(\mathcal{B}_i, \mathcal{B}_j) = \text{AUROC}(g_{\mathcal{B}_i} \mid \mathcal{B}_j), \quad j \neq i. \quad (3)$$

A large gap $\tau^{\text{within}} \gg \tau^{\text{cross}}$ indicates that the surface-form gap is benchmark-specific: produced by particular choices of generators, prompts, and human sources, rather than reflecting a stable property of machine generation. Specifically, we repeat this experiment with L1 and L2 logistic regression across four regularization strengths, and ridge classification to rule out classifier choice as a source of transfer failure.

Detector sensitivity to surface form. Let h be a detector that assigns a score $s^h(x) \in \mathbb{R}$ to each text. We measure the sensitivity of s^h to surface form in three complementary ways.

First, we regress s^h on ϕ (ridge, 5-fold CV) to quantify how much of the detector’s score is predictable from surface form, reporting the coefficient of determination R^{2h} and the residual AUROC. The measurement is correlational: consistent with h using surface form and with a common cause (benchmark construction) shaping both.

Second, we compare each detector’s AUROC on the full AI set to a subset of AI samples matching the human median on at least two of the three strongest AI-direction features (paragraph count, hapax ratio, paragraph-length variability) to measure sensitivity to co-occurring surface differences, without holding content or generator fixed.

Third, we intervene on surface form directly. An intervention $\pi : x \mapsto \pi(x)$ merges paragraph breaks in a text down to the human median, leaving every word unchanged. For each detector we report the AUROC shift

$$\Delta\text{AUROC}^h = \text{AUROC}(s^h) - \text{AUROC}(s^h \circ \pi), \quad (4)$$

and the fraction of the human/AI mean-score gap closed by π . We apply π to the five likelihood-based detectors (Log-Rank, Log-Likelihood, Rank, Entropy, LRR), which all aggregate token log-probabilities from the same underlying model.

4 Experimental Setup

Benchmarks. We use RAID (Dugan et al., 2024) as our primary benchmark for its scale and generator diversity: 26,742 samples across 8 domains (news, books, poetry, recipes, Reddit, reviews, abstracts,

Table 1: Detector AUROC across three benchmarks (mean over 5 seeds). **Surface-form** (highlighted) is the surface-form diagnostic; it reaches the top score on every benchmark.

Method	RAID	M4	NYT
Surface-form	0.974	0.968	0.999
RADAR	0.889	0.715	0.828
RoBERTa	0.846	0.810	0.795
Binoculars	0.850	0.899	0.998
LRR	0.811	0.709	0.974
Fast-DetectGPT	0.794	0.834	0.988
DivEye	0.785	0.850	0.936
Rank	0.782	0.742	0.963
Log-Rank	0.778	0.830	0.935
Log-Likelihood	0.762	0.828	0.899
Entropy	0.671	0.720	0.588

Wikipedia) from 11 generators, sampled into class-balanced subsets (50% human, 50% AI) averaged over 5 seeds. To test generalization, we replicate on M4 (Wang et al., 2024) (4 domains, 6 generators) and NYT (Roy et al., 2025) (New York Times articles paired with recent LLMs); sampling details in Appendix B.

Features. We extract 86 features that describe surface properties of each text, covering document structure, length, lexical diversity, punctuation, shallow style, structural patterns, and shallow syntax. Appendix C lists all features. All main results use the full 86-feature *surface-form* set. In §5.1 we additionally consider a *strict formatting* subset (document structure, length, punctuation, structural patterns, and character-ratio features only, excluding the dedicated lexical, stylistic, and syntactic feature families) to show that the classifier’s performance does not depend on the more content-adjacent feature families.

Detectors. We evaluate eight zero-shot detectors across the main families: Binoculars (Hans et al., 2024) (perplexity ratios), DivEye (Basani and Chen, 2025) (surprisal diversity), Fast-DetectGPT (Bao et al., 2024) (curvature), and the likelihood-based log-likelihood, log-rank, rank, entropy (Gehrmann et al., 2019), and LRR (Su et al., 2023). We also evaluate two trained detectors, the off-the-shelf RoBERTa classifier (Solaiman et al., 2019) and the adversarially trained RADAR (Hu et al., 2023), to test whether the confound is specific to zero-shot methods. Details in Appendix E.

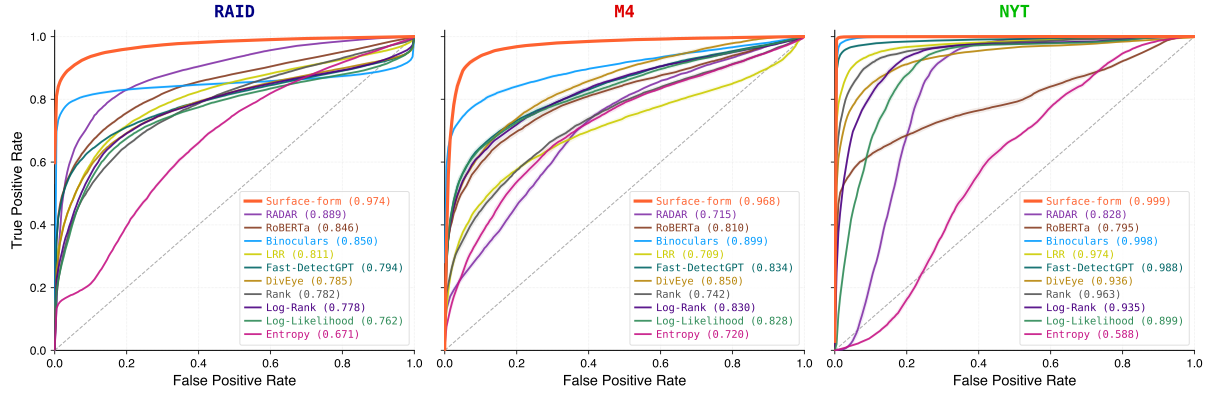


Figure 1: ROC curves on RAID, M4, and NYT (mean across 5 seeds). **Surface-form** (orange) sits above every detector on every benchmark.

5 Results

5.1 Surface form separates human and machine text within benchmarks

Surface-form reaches 0.974 AUROC on **RAID** (Table 1), above the eight zero-shot detectors and both supervised detectors. The pattern holds on **M4** (0.968) and **NYT** (0.999).

This does not depend on lexical, stylistic, or syntactic features. The strict formatting feature set ϕ^{strict} (structural, length, punctuation, and character-ratio features only) reaches within 0.012 AUROC of the full 86-feature ϕ on **RAID**, and within 0.012 on all three benchmarks (Figure 3). Document structure alone (nine paragraphing and whitespace features) reaches 0.932 on **RAID**, above every detector. A length-only baseline reaches only 0.624 on **RAID**, ruling out length as the primary factor. On **NYT**, length alone reaches 0.934, reflecting benchmark construction: generations are prompted from article abstracts, so human and AI text differ in length by design. The largest AI-direction coefficients on **RAID** are paragraph count, hapax ratio, and paragraph-length variability (Figure 4); AI text uses more paragraphs than human text.

5.2 Patterns are benchmark-specific

The classifier transfers across domains but not across benchmarks. Within-benchmark transfer is high: $\tau^{\text{within}}(\text{RAID}) = 0.945$, with all held-out **RAID** domains above 0.86 (Table 8). However, cross-benchmark transfer collapses: $\tau^{\text{cross}}(\text{RAID}, \text{M4}) = 0.28$ (Figure 2). The same asymmetry $\tau^{\text{within}} \gg \tau^{\text{cross}}$ holds for all three benchmarks. Each benchmark contains a surface-form gap, but the gap sits in different features, since each uses different generators and prompts.

The transfer failure is not an artifact of classifier choice. We repeat the cross-benchmark experiment with L1 and L2 logistic regression at four regularization strengths ($C \in \{0.01, 0.1, 1.0, 10.0\}$), and ridge classification. All variants show the same asymmetry: within-**RAID** transfer stays near 0.945, and **RAID** $\rightarrow$ **M4** transfer stays near 0.28.

5.3 Detector scores are associated with surface form to varying degrees

Surface form predicts a substantial fraction of detector-score variance. Across detectors, surface form explains up to 69% of detector-score variance on **RAID**, 60% on **M4**, and 77% on **NYT** (see Appendix D). The residual AUROC falls to 0.60 to 0.73 on **RAID**, but detectors retain some discriminative signal after regressing out surface form.

Detectors deteriorate on AI text with human-like surface form. On the subset of AI samples matching the human median on at least two of the three strongest AI-direction features, zero-shot detectors drop by up to 20% (**Log-Likelihood** 20.3%, **Log-Rank** 19.4%; Table 6). The two supervised detectors drop less (**RADAR** 3.5%, **RoBERTa** 6.5%).

Paragraph-normalisation shifts some detector scores but not others. Under π , $\Delta\text{AUROC}^{\text{Entropy}} = 0.024$ (from 0.673 to 0.649) and $\Delta\text{AUROC}^{\text{LRR}} = 0.030$ (from 0.842 to 0.812), closing 12.4% and 17.2% of the human/AI mean-score gap over 5 seeds (Table 2). $\Delta\text{AUROC}^{\text{Log-Rank}}$ is small; $\Delta\text{AUROC}^{\text{Log-Likelihood}}$ and $\Delta\text{AUROC}^{\text{Rank}}$ are essentially zero. About a third of AI samples in **RAID** are already single-paragraph and untouched by the intervention, so the effect on modified samples is larger. Paragraph structure therefore has a direct

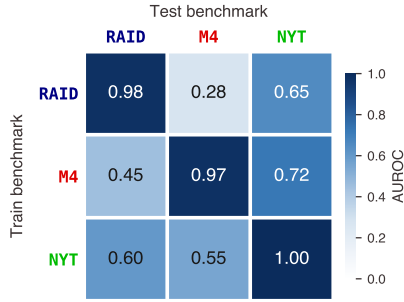


Figure 2: Cross-dataset transfer of the surface-form classifier (AUROC, mean across 5 seeds). The bright diagonal shows strong within-dataset performance; off-diagonal performance shows that the learned surface-form patterns transfer inconsistently across benchmarks.

Table 2: Effect of paragraph-normalization on detector AUROC (RAID, mean over 5 seeds). **Orig**: AUROC before intervention. **Norm**: AUROC after. **Δ gap**: fraction of the human/AI mean-score gap closed. Bolded values mark meaningful shifts.

Detector	Orig	Norm	Δ gap
LRR	0.842	0.812	17.2% $\pm$ 1.2
Entropy	0.673	0.649	12.4% $\pm$ 1.1
Log-Rank	0.778	0.770	2.1% $\pm$ 0.8
Log-Likelihood	0.762	0.761	≈ 0
Rank	0.782	0.782	≈ 0

causal effect on the scores of two of the five detectors we test.

6 Conclusion

We find an important and underexamined confound in AI-text detection evaluation: current benchmarks contain strong surface-form signals that distinguish human from machine-generated text within individual benchmarks, while these signals vary substantially across benchmarks.

Across RAID, M4, and NYT, a surface-form classifier reaches or exceeds the AUROC of every detector we evaluate, and the strict formatting subset ϕ^{strict} reaches within 0.012 AUROC of the full 86-feature set on all three benchmarks. The learned patterns transfer within a benchmark (0.945) but not across benchmarks (0.28), indicating that the surface-form gap reflects benchmark construction rather than a stable property of machine-generated text across evaluation settings. Further, detector scores are substantially predictable from surface form, and a paragraph-normalisation intervention causally shifts scores for two of five likelihood-based detectors we test.

Together, these findings are consistent with a

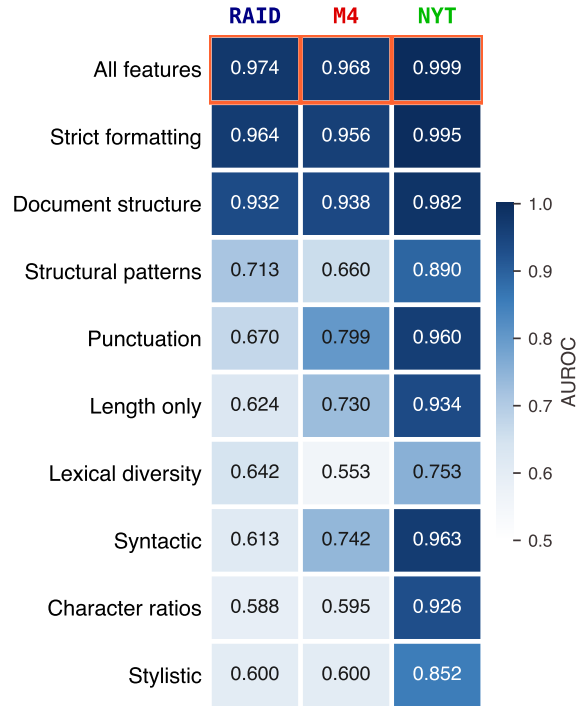


Figure 3: Feature-family ablation on RAID, M4, and NYT. Structural surface features (top three, shaded) dominate; lexical, stylistic, and syntactic features contribute little. Length alone is not the driver.

Clever Hans effect, where high benchmark performance does not by itself imply that a detector has learned a stable and transferable signal of machine generation. A simple surface-feature classifier can help show this effect. When surface features approach detector performance, the benchmark contains exploitable surface cues. We therefore recommend reporting surface-feature separability alongside detector performance.

Limitations

We study three English-language benchmarks, and surface conventions and benchmark construction may differ across languages and datasets. Our cross-benchmark analysis shows that the recovered surface patterns are not stable across benchmarks, but does not isolate whether these differences arise from human sources, generators, prompts, domains, or their interaction. Our diagnostic uses a finite set of hand-crafted features and linear models, and may therefore miss other or nonlinear surface cues. Finally, our causal evidence is limited to paragraph normalisation for five likelihood-based detectors, so it establishes sensitivity to paragraph structure rather than to surface form in general.

Ethical Considerations. Our work reveals vulnerabilities in AI-generated text detectors, which could potentially be exploited to evade detection. However, we believe disclosure is appropriate: the formatting differences we identify are already present in public benchmarks, and understanding detector limitations is necessary for improving them. We do not release evasion tools or provide specific guidance for circumventing detectors. Our findings highlight the need for more robust evaluation practices rather than enabling misuse. The detectors we evaluate are publicly available, and our analysis uses only published benchmark datasets with no personal or sensitive data.

Acknowledgments

The work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding. CEB acknowledges her AI4S fellowship within the “Generación D” initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.

References

- Yassir El Attar, Esra Dönmez, Maximilian Maurer, and Agnieszka Falenska. 2026. A Systematic Analysis of Linguistic Features in AI-Generated Text Detection Across Domains and Models. *arXiv preprint arXiv:2606.04177*.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.
- Advik Raj Basani and Pin-Yu Chen. 2025. Diversity boosts AI-generated text detection. *arXiv preprint arXiv:2509.18880*.
- Angana Borah, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2023. [Measuring spurious correlation in classification: “clever hans” in translationese](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 196–206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Claudio Borile and Carlo Abrate. 2025. How to generalize the detection of ai-generated text: Confounding neurons. In *EMNLP (Findings)*, pages 25461–25476.
- Koel Dutta Chowdhury, Richa Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492.
- Lang Gao, Xuhui Li, Chenxi Wang, Mingzhe Li, Wei Liu, Zirui Song, Jinghui Zhang, Rui Yan, Preslav Nakov, and Xiuying Chen. 2026. When personalization tricks detectors: The feature-inversion trap in machine-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43143–43171.
- 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.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. In *International Conference on Machine Learning*, pages 17519–17537. PMLR.
- Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. 2025. Learning to rewrite: Generalized llm-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 6421–6434.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36.
- 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.
- 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. *Advances in Neural Information Processing Systems*, 36:27469–27500.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.
- Andrea Pedrotti, Michele Papucci, Cristiano Ciacio, Alessio Miaschi, Giovanni Puccetti, Felice Dell’Orletta, and Andrea Esuli. 2025. Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3010–3031.
- Oskar Pfungst. 1911. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Xiao Pu, Zepeng Cheng, Lin Yuan, Yu Wu, and Xiuli Bi. 2026. [Breaking the generator barrier: Disentangled representation for generalizable AI-text detection](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2586–2598, San Diego, California, United States. Association for Computational Linguistics.
- Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef Van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611.
- Rajarshi Roy, Nasrin Imanpour, Ashhar Aziz, Shashwat Bajpai, Gurpreet Singh, Shwetangshu Biswas, Kapil Wanaskar, Parth Patwa, Subhankar Ghosh, Shreyas Dixit, and others. 2025. A Comprehensive Dataset for Human vs. AI Generated Text Detection. *arXiv preprint arXiv:2510.22874*.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412.
- Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and others. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.
- Ruochong Xiong, Qien Li, Wangwang Lian, Yulong Wan, Hanlin Xue, Zhouxing Tan, Han Yang, Fengyu Lu, and Junfei Liu. 2026. Verifiable llm-generated text detection via projected semantic-structural distributions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14005–14042.
- Xiao Yu, Yi Yu, Dongrui Liu, Kejiang Chen, Weiming Zhang, Nenghai Yu, and Jing Shao. 2025. [EvoBench: Towards real-world LLM-generated text detection benchmarking for evolving large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14605–14620, Vienna, Austria. Association for Computational Linguistics.

Sergio E Zanotto and Segun Aroyehun. 2025. Linguistic and embedding-based profiling of texts generated by humans and large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

A Related Work

Zero-shot detectors measure probability-based signatures of machine generation. DetectGPT (Mitchell et al., 2023) identifies text at local maxima of the log probability function; Fast-DetectGPT (Bao et al., 2024) approximates curvature through efficient sampling; Binoculars (Hans et al., 2024) computes perplexity ratios between LLM pairs. These methods report strong benchmark results without requiring training data. Trained approaches like Ghostbuster (Verma et al., 2024) learn classifiers over model-derived features and explicitly evaluate generalization across domains, prompts, and language models. Prior work has documented brittleness across both paradigms: detectors degrade 20–32% under changes in decoding strategy or repetition penalty (Dugan et al., 2024), and directly trained classifiers can struggle to generalize to unseen domains (Hao et al., 2025). Paraphrasing attacks evade detection entirely (Sadasivan et al., 2023; Krishna et al., 2023).

Prior attacks modify AI-generated text to evade detection; we show that formatting differences in baseline benchmark data already inflate accuracy before any manipulation. This connects to shortcut learning in NLP more broadly. In natural language inference, hypothesis-only baselines revealed that models exploit annotation artifacts rather than learning entailment (Poliak et al., 2018; Gururangan et al., 2018). We apply analogous methodology to AI detection: a surface-form baseline that tests whether benchmarks contain exploitable surface confounds. Borah et al. (2023) show that named entities can function as confounds in distinguishing original from translated texts in the same language. Similar concerns have been raised in the study of translationese, where hand-crafted linguistic features and neural representations both distinguish translated from original text (Pylypenko et al., 2021), but where the distinction can reflect construction-specific artefacts of the translation process rather than a stable linguistic signature (Chowdhury et al., 2022). The distinction between human and machine text can therefore contain both properties that generalize across settings and properties that are specific to a particular

evaluation distribution. Work on linguistic profiling shows that deep features (syntactic structure, semantic similarity) achieve only limited accuracy distinguishing human from machine text (Zanotto and Aroyehun, 2025), and recent analysis finds that many linguistic indicators of AI-generated text are strongly dependent on the domain and model in which they are measured (Attar et al., 2026).

B Dataset Details

RAID (Dugan et al., 2024) contains human and AI-generated text across 8 domains (news, books, poetry, recipes, Reddit, reviews, abstracts, Wikipedia) from 11 language models. We exclude code as formatting features differ substantially. Human samples (13,371) constitute the limiting factor; we sample an equal number of AI-generated texts to create class-balanced subsets (50% human, 50% AI) of 26,742 samples per seed, averaged across 5 seeds. We sample balanced subsets for each seed:

Table 3: **RAID** sampling statistics per seed.

Total samples	26,742
Human samples	13,371
AI samples	13,371
Domains	8
AI models	11
Seeds	5

Domains: News, Books, Poetry, Recipes, Reddit, Reviews, Abstracts, Wikipedia.

AI Models: ChatGPT, GPT-4, GPT-3, Cohere, LLaMA, MPT, Mistral, and others.

M4 (Wang et al., 2024) is a multilingual, multi-generator benchmark. We use the English portion: 4 domains (arXiv, Wikipedia, WikiHow, Reddit) with text from 6 generators (GPT-3.5, ChatGPT, GPT-4, Cohere, Dolly-v2, BLOOMz). We sample 1,000 paired human-AI texts per domain to create class-balanced subsets (50% human, 50% AI) of 8,000 samples per seed, averaged across 5 seeds. We sample balanced subsets for each seed:

Table 4: **M4** sampling statistics per seed.

Samples per domain	2,000 (1,000 pairs)
Total samples	8,000
Domains	4
AI models	6
Seeds	5

Domains: arXiv, Wikipedia, WikiHow, Reddit.

AI Models: GPT-3.5 (davinci-003), ChatGPT, GPT-4, Cohere, Dolly-v2, BLOOMz. For each

human text, we randomly select one of its machine generations, ensuring model diversity across the dataset.

NYT (Roy et al., 2025) is a comprehensive dataset comprising over 58,000 text samples that combine authentic New York Times articles with synthetic versions generated by state-of-the-art LLMs (Gemma-2-9b, Mistral-7B, Qwen-2-72B, LLaMA-3-8B, Yi-Large, and GPT-4). The dataset uses original article abstracts as prompts for generation. While RAID and M4 use older models (GPT-3, ChatGPT, Cohere), NYT includes 2024-era RLHF-tuned generators, testing whether surface-form confounds persist with recent models. We sample balanced subsets for each seed:

Table 5: **NYT** sampling statistics per seed.

Total samples	8,000
Human samples	4,000
AI samples	4,000
Domains	1 (News)
AI models	6
Seeds	5

Source: New York Times articles spanning January 2000 to present.

AI Models: Gemma-2-9b, Mistral-7B, Qwen-2-72B, LLaMA-3-8B, Yi-Large, GPT-4o.

C Surface-Form Features

We extract surface-level features organized into 8 categories:

C.1 Document Structure

- `n_paragraphs`: Number of paragraphs (double newlines + 1)
- `n_newlines`: Total newline count
- `n_bullets`: Bullet point count
- `n_numbered_lists`: Numbered list items
- `n_headers`: Header-like lines (ending with colon)
- `has_code_block`: Presence of code formatting
- `whitespace_ratio`: Whitespace characters / total characters
- `avg_paragraph_len`: Mean paragraph length (characters)
- `paragraph_len_std`: Paragraph length standard deviation

C.2 Character Ratios

Character-level surface ratios independent of tokenization:

- `uppercase_ratio`: uppercase characters / total characters
- `digit_ratio`: numeric characters / total characters
- `special_char_ratio`: non-alphanumeric, non-whitespace characters / total characters

C.3 Length Statistics

- `char_len`, `word_len`, `n_sentences`: Document length measures
- `avg_word_len`, `word_len_std`: Word length statistics
- `avg_sentence_len`, `sentence_len_std`: Sentence length statistics
- `max_sentence_len`, `min_sentence_len`: Sentence length extremes

C.4 Lexical Features

- `vocab_richness`: Unique words / total words
- `hapax_ratio`: Words appearing once / total words
- `long_word_ratio`: Words >10 characters
- `short_word_ratio`: Words ≤3 characters
- `avg_syllables`: Estimated syllables per word
- `flesch_kincaid`: Readability score

C.5 Punctuation

Frequency (per word) of: periods, commas, semicolons, colons, exclamation marks, question marks, ellipses, dashes, parentheses (open/close), brackets, quotation marks (single/double), apostrophes, slashes.

C.6 Stylistic Features

- `transition_ratio`: Transition words (however, therefore, etc.)
- `hedge_ratio`: Hedge words (maybe, perhaps, etc.)
- `formal_ratio`: Formal vocabulary (utilize, demonstrate, etc.)
- `contraction_ratio`: Contractions (don't, can't, etc.)
- `first/second/third_person_ratio`: Pronoun usage
- `passive_indicator`: Passive voice patterns

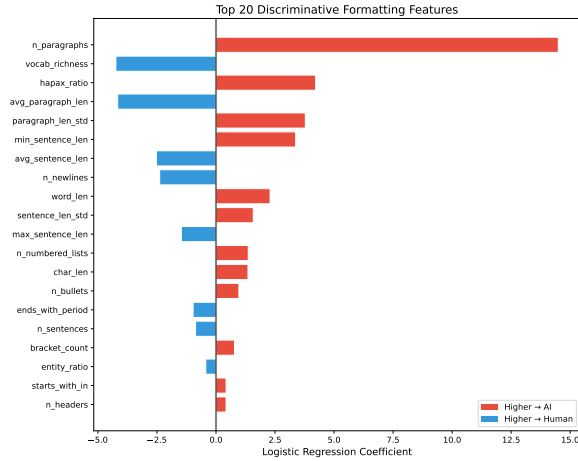


Figure 4: Top 20 features by logistic regression coefficient. Red: higher values predict AI; blue: higher values predict human.

C.7 Structural Patterns

- `starts_with_*`: Binary flags for 18 common opening words/phrases (in, the, this, it, there, here, as, with, when, while, although, because, since, if, for, to, a, an)
- `ends_with_period/question/exclamation`: Ending punctuation
- `has_url`, `has_email`, `has_hashtag`, `has_mention`: Special patterns

C.8 Syntactic Features

Extracted using spaCy:

- `avg_dep_depth`, `max_dep_depth`: Dependency tree depth
- `pos_*_ratio`: Part-of-speech ratios (noun, verb, adj, adv, pron, det, adp, conj, punct)
- `entity_ratio`: Named entity density

D Which surface-form features discriminate?

Figure 4 shows the top 20 discriminative surface-form features. Paragraph count dominates (coefficient +14.5): AI-generated text uses significantly more paragraphs than human text. Other AI-indicative features include hapax ratio and paragraph length variability; human text exhibits richer vocabulary and longer paragraphs.

D.1 RAID Detailed Results

We rank surface-form features by logistic regression coefficient magnitude and use the three strongest AI-direction features (paragraph count,

hapax ratio, paragraph length variability) to compute an “AI-score” (0–3) for each AI sample, counting how many features exceed the human median. AI samples scoring ≤ 1 are classified as “human-like” (i.e., human-typical on at least two of three features). In Table 6, we compare detector performance on the full dataset versus a subset containing all human samples plus only human-like AI samples. Zero-shot detectors drop by up to 20% on AI text with human-like surface form; the supervised detectors drop less. Residual AUROCs (0.60 to 0.73) stay above chance, indicating signal beyond surface form.

Table 6: Detailed detector results on RAID (mean over 5 seeds). **AUROC**: full-set AUROC. **H-like**: AUROC on the AI subset with human-like surface form values. **Drop**: relative decrease from AUROC to H-like. **Residual**: AUROC after regressing detector scores on surface-form features. **R²**: variance in detector scores explained by surface form.

	AUROC	H-like	Drop	Residual	R ²
Surface-form	0.974	0.924	5.3%	0.500	1.00
RADAR	0.889	0.858	3.5%	0.679	0.46
RoBERTa	0.846	0.791	6.5%	0.659	0.38
Binoculars	0.850	0.711	16.4%	0.729	0.61
LRR	0.811	0.686	15.4%	0.654	0.57
Fast-DetectGPT	0.794	0.670	15.6%	0.601	0.45
DivEye	0.785	0.762	2.9%	0.682	0.51
Rank	0.782	0.685	12.4%	0.628	0.18
Log-Rank	0.778	0.628	19.4%	0.677	0.65
Log-Likelihood	0.762	0.607	20.3%	0.674	0.69
Entropy	0.671	0.568	15.4%	0.622	0.66

D.2 Residual Analysis on RAID

Table 7 shows residual performance after regressing out surface-form features. The Surface-form drops to 0.5 (chance), confirming complete signal removal. All detectors retain above-chance residual AUROCs (0.60–0.73), indicating signal beyond surface-form features.

D.3 Cross-Domain Transfer (within RAID)

Table 8 shows that the learned surface-form patterns transfer strongly across RAID domains.

D.4 Cross-Benchmark Transfer

Table 9 shows that transfer is substantially weaker and asymmetric across benchmarks.

Table 7: Residual analysis on RAID (mean over 5 seeds). **AUROC**: full-set AUROC. **Residual**: AUROC after regressing detector scores on surface-form features. **R²**: variance in detector scores explained by surface form.

	AUROC	Residual	R ²
Surface-form	0.974	0.500	1.00
RADAR	0.889	0.679	0.46
RoBERTa	0.846	0.659	0.38
Binoculars	0.850	0.729	0.61
LRR	0.811	0.654	0.57
Fast-DetectGPT	0.794	0.601	0.45
DivEye	0.785	0.682	0.51
Rank	0.782	0.628	0.18
Log-Rank	0.778	0.677	0.65
Log-Likelihood	0.762	0.674	0.69
Entropy	0.671	0.622	0.66

Table 8: Cross-domain transfer within RAID (train on 7 domains, test on held-out).

Held-out Domain	AUROC	N
Recipes	0.992±0.002	3,544
News	0.979±0.002	3,560
Wiki	0.973±0.001	3,558
Books	0.967±0.003	3,562
Poetry	0.953±0.005	3,542
Reviews	0.933±0.007	1,886
Abstracts	0.900±0.016	3,532
Reddit	0.866±0.006	3,558
Mean	0.945±0.041	

D.5 Residual Analysis on M4

Table 10 shows results on M4. Surface-form (0.968 AUROC) exceeds all detectors, with Binoculars highest at 0.899. Surface form explains up to 60% of detector variance (R^2). Residual AUROCs (0.54–0.75) indicate detectors retain signal beyond surface cues.

D.6 Residual Analysis on NYT

Table 11 shows results on NYT. The surface-form confound strengthens: formatting achieves 0.999 AUROC and explains up to 77% of variance, higher than RAID (69%). Residual AUROCs drop to 0.53–0.68, indicating stronger predictability from surface form on NYT.

D.7 Implementation Details

Preprocessing and truncation. All detector inputs are truncated to 512 tokens using the detector’s own tokenizer. For the surface-form classifier, features are extracted from the untruncated text (surface statistics are cheap to compute over the

Table 9: Cross-benchmark transfer of the surface-form classifier (mean ± std over 5 seeds). Rows are the training benchmark; columns are the evaluation benchmark.

Train	RAID	M4	NYT
RAID	0.974 ± 0.001	0.281 ± 0.004	0.65 ± 0.007
M4	0.448 ± 0.009	0.968 ± 0.002	0.72 ± 0.012
NYT	0.643 ± 0.009	0.545 ± 0.021	1.000 ± 0.000

Table 10: M4 results (mean over 5 seeds).

	AUROC	Residual	R ²
Surface-form	0.968	0.500	1.00
Binoculars	0.899	0.753	0.48
DivEye	0.850	0.687	0.49
Fast-DetectGPT	0.834	0.743	0.58
Log-Rank	0.830	0.742	0.54
Log-Likelihood	0.828	0.743	0.60
Rank	0.742	0.648	0.30
Entropy	0.720	0.664	0.54
LRR	0.709	0.536	0.56

full document).

Fold construction. The surface-form classifier uses 5-fold stratified cross-validation, stratified by label (human/AI). Folds are fixed per seed across all methods so that direct comparisons between the classifier and detectors use identical evaluation subsets. Standard scaling is fit only on training folds.

Seeds. Dataset-sampling seeds are 42, 123, 456, 789, and 1011. All main results are reported as means over these 5 seeds; where a standard deviation is given, it is computed across seeds.

Score direction validation. Each detector’s score-direction (higher-score = machine or higher-score = human) follows the method’s documented polarity (see Appendix E), and is verified by checking that AUROC > 0.5 on the full RAID set for each seed. Any detector whose orientation was ambiguous was validated against its original paper’s reference implementation.

Pairing (M4). For M4, one machine generation is sampled per human text (of the six generators available for each human source), keeping the human/AI comparison paired within each seed. The 6 generators are approximately balanced across seeds.

Detector implementations. Binoculars uses the official implementation with Falcon-7B and

Table 11: NYT results (mean over 5 seeds).

	AUROC	Residual	R ²
Surface-form	0.999	0.500	1.00
Binoculars	0.998	0.684	0.77
Fast-DetectGPT	0.988	0.604	0.68
Rank	0.963	0.533	0.45
DivEye	0.936	0.612	0.60
Log-Rank	0.935	0.626	0.74
Log-Likelihood	0.899	0.622	0.73
Entropy	0.588	0.543	0.58

Falcon-7B-Instruct as observer and performer, default (low-fpr) scoring mode. Fast-DetectGPT uses the official implementation with GPT-Neo-2.7B as the scoring model and GPT-J-6B as the reference model. DivEye computes token-level surprisal using GPT-2 and extracts distributional statistics (mean, variance, skewness), first-order differences, and second-order dynamics. Likelihood-based methods (Log-Likelihood, Log-Rank, Rank, Entropy, LRR) are computed using GPT-2 XL following Su et al. (2023).

Surface-form Classifier We train a logistic regression classifier on the 86 surface-form features described in Appendix C. Implementation details:

- **Model:** Logistic regression with L2 regularization (scikit-learn LogisticRegression)
- **Regularization:** $C = 1.0$ (default)
- **Evaluation:** 5-fold stratified cross-validation
- **Feature preprocessing:** Standard scaling (zero mean, unit variance)
- **Seeds:** 5 random seeds (42, 123, 456, 789, 1011) for dataset sampling; results averaged

Compute. All detector inference was run on a single NVIDIA A100 (80GB), total ~96 GPU-hours across all seeds and benchmarks.

E Detection Methods

We evaluate zero-shot detectors representing different approaches to measuring probability-based signatures, plus two supervised detectors.

Perplexity-based. Binoculars (Hans et al., 2024) computes perplexity divided by cross-perplexity using two LLMs (Falcon-7B and Falcon-7B-Instruct). The ratio normalizes for prompt-induced perplexity variations. Lower scores indicate machine text.

Diversity-based. DivEye (Basani and Chen, 2025) extracts statistical features from token-level surprisal patterns, capturing how unpredictability fluctuates across text. Human text exhibits greater surprisal diversity than LLM outputs; lower diversity indicates machine text.

Curvature-based. Fast-DetectGPT (Bao et al., 2024) approximates the DetectGPT (Mitchell et al., 2023) probability curvature without expensive perturbation sampling. Higher curvature indicates machine text.

Likelihood-based. Log-Likelihood, Log-Rank, Rank, and Entropy (Gehrmann et al., 2019), along with LRR are computed using GPT-2 XL (1.5B parameters) following Su et al. (2023). For each text, we compute token-level log probabilities and aggregate according to each method’s specification.

Supervised. RADAR (Hu et al., 2023) is a RoBERTa-based classifier trained adversarially with a paraphraser. RoBERTa is the OpenAI GPT-2 output detector (Solaiman et al., 2019), a RoBERTa-large model trained on human and GPT-2 generations.

We evaluate zero-shot detectors representing different approaches to measuring probability-based signatures:

F Use of AI Assistants

AI assistants were used for editing prose and debugging code. All scientific claims, experimental design, and analysis were conducted by the authors.