Ranking Human and LLM Texts Using Locality Statistics

Yiyang Wang, Chen Ding, Hangfeng He
Department of Computer Science, University of Rochester
{yiyang.wang, hangfeng.he}@rochester.edu, cding@cs.rochester.edu

Abstract

The paper extends the Data Movement Distance (DMD) – a metric defined to measure the locality in computer memory – to text by defining a normalized version called nDMD. A key feature of nDMD is a new term designed to better characterize low-frequency tokens.

By evaluating nDMD on English subset of the M4 dataset and GenAI detection shared task, the paper shows three key findings. First, nDMD is systematically higher in human-written text than in machine-generated text. Second, nDMD-based features not only outperform frequency baselines but also improve overall performance when combined. Finally, the proposed DMD normalization is more effective in distinguishing human and machine text than alternative normalization approaches.

1 Introduction

The rapid advancement of large language models (LLMs) has made machine-generated text (MGT) increasingly fluent and difficult to distinguish from human writing (OpenAI et al., 2024; Mitchell et al., 2023). Reliable detection of MGT has therefore become critical not only in high-stakes domains such as medicine, education, and law, but also in addressing public concerns surrounding misinformation and malicious misuse (Wu et al., 2025; Weber-Wulff et al., 2023). While the detection task predates LLMs (Badaskar et al., 2008), recent approaches often rely on transformer architectures themselves (Wu et al., 2025), leveraging fine-tuned classifiers (Ippolito et al., 2020), model-derived statistics (Mitchell et al., 2023; Gehrmann et al., 2019), or prompt-based discriminators (Koike et al., 2024; Bhattacharjee and Liu, 2024).

While transformer-based detection methods achieve impressive accuracy, they often demand substantial resources for training and inference, and their decision-making processes remain largely opaque (Wu et al., 2025). By contrast, classical statistical features directly collected from the text, such as n-gram frequencies (Badaskar et al., 2008; McGovern et al., 2025) and entropy (Lavergne et al., 2008), offer complementary strengths. Though generally less competitive in raw accuracy (Mindner et al., 2023), the statistical methods are attractive in settings where their lightweight nature, efficiency, and interpretability are valued. In this paper, we extend this line of work by proposing *locality statistics* as an alternative and complementary class of features for MGT detection.

Locality statistics measure how items are *reused* across a sequence rather than how often they appear. We focus on two key measures: **Reuse Distance (RD)** (Zhong et al., 2009), or synonymously Least Recently Used (LRU) Stack Distance (Mattson et al., 1970), and **normalized Data Movement Distance (nDMD)**, a text-specialized extension of the Data Movement Distance (Smith et al., 2022), which together capture the spacing and cumulative effort of token reuse. Originally developed in computer systems, RD and DMD quantify the machine-independent cache efficiency of workloads under the LRU cache replacement policy. In text, however, locality reflects the generative process itself: human writing embodies cognitive constraints on memory and discourse planning, whereas LLM-generated text reflects the inductive biases of model architectures and decoding heuristics. Thus, locality statistics act as *fingerprints* of the generator, offering a distinctive signal for distinguishing human and machine text.

We test this idea using the English subset of the M4 dataset (Wang et al., 2024), which contains paired human-written and model-generated text across domains and across LLM families. Our study has three main findings. First, we characterize human vs. machine text and show that human text exhibits consistently higher nDMD and

Findings of the Association for Computational Linguistics: EACL 2026, pages 5337–5348
March 24-29, 2026 ©2026 Association for Computational Linguistics

larger variance, indicating richer and more varied reuse patterns. Second, we evaluate variants of RD treatment and show that our definition of nDMD yields the strongest class separation. Third, in classification tasks, locality features alone rival the frequency-based baselines of bag-of-words, while their combination achieves the highest overall accuracy. Our contributions are twofold:

- **Adapting locality metrics to text.** We extend the *Data Movement Distance* (DMD) metric from memory-system analysis to natural language, defining a normalized variant (nDMD) that captures token reuse patterns while accounting for low-frequency tokens.
- **Empirical validation.** We use English M4 (Wang et al., 2024) and *GenAIDetect* (Wang et al., 2025) datasets to show that nDMD-based features both outperform and complement frequency baselines in discriminating between human and machine text in both controlled and open-domain settings.

2 Locality Metrics for Natural Languages

We develop and analyze metrics that capture locality patterns in text, drawing an analogy to memory locality in computer systems. In systems, a workload is a sequence of memory accesses, and locality statistics measure cache efficiency independent of hardware. Similarly, we treat a text document as a sequence of tokens and compute locality statistics over them. As a fundamental statistical unit in modern NLP, a token is analogous to a memory reference, and locality statistics quantify how closely these past "references" are reused.

Our primary measurement is **Reuse Distance (RD)**[1], defined as *number of distinct tokens from a token's previous occurrence up to (and including) its current one*. Every repeated token therefore has an RD determined by the distinct intervening tokens since its last appearance, while the first occurrence of each token lacks a well-defined RD. In the context of a fully associative LRU cache, RD quantifies a local working-set size, the minimum cache capacity required to keep a token readily available for reuse.

Building on RD, **Data Movement Distance (DMD)** (Smith et al., 2022) summarizes the total data movement by summing the square-root RDs:

$$\mathrm{DMD}(D) = \sum_{d \in D} \sqrt{d},$$

where D contains all observed RDs in the reference stream. The square-root operation reflects the two-dimensional layout of computer memory. Whether the same principle applies to human memory is not yet established. However, empirically, this treatment yielded the best separation between human and machine-generated text in our experiments (Section 4.2).

DMD was developed for asymptotic analysis, where the frequency of data reuse increases at a faster asymptotic rate than the number of unique data items as the input size grows (Smith et al., 2022; Ding et al., 2023). In its original formulation, only data reuses are counted, and the first-time references are ignored in the measure of DMD. Word frequencies, however, are highly skewed, and many tokens may not observe any reuses within a document. To make the measure comparable across documents and account for sparsely reused tokens, we define a **normalized DMD (nDMD)**:

$$\mathrm{nDMD}(D, A) = \frac{\mathrm{DMD}(D) + \sum_{i=1}^{|A|} \sqrt{i}}{|D| + |A|}. \quad (1)$$

Beyond mean normalization, we introduce an additional parameter A, the set of unique tokens in the text, which accumulates incremental imaginary reuse distances (I-RDs) for first token occurrences. This is equivalent to prepending the unique tokens in reverse order (Figure 1). I-RDs represent a universal best-case RD for first references, i.e., the smallest reuse distances these tokens could attain given a fixed working set (vocabulary). This normalization removes document-length effects while softly incorporating working-set size (A), a memory constraint not captured by raw RD, particularly when reuses are sparse.

It is worth stressing that, even after normalization, nDMD often increases sublinearly with document length. Longer texts typically introduce more distinct tokens (i.e., larger vocabularies), which raises the maximum reuse distance and accumulates more imaginary reuse distances (I-RDs) from first occurrences. Consequently, higher nDMD reflects a larger effective working set, manifested both in richer local reuse patterns and in the overall vocabulary of the text.

[1] It is first defined by Mattson et al. (1970) as the LRU Stack Distance and later commonly known as the Reuse Distance (Zhong et al., 2009; Yuan et al., 2019). The asymptotic cost in RD measurement has been reduced in time using a tree (Olken, 1981; Ding and Zhong, 2003) and in space (Wires et al., 2014).

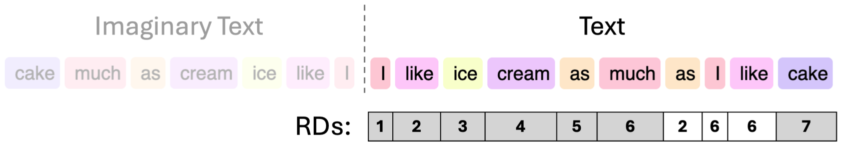


Figure 1: Example Reuse Distances (RDs) for text. Shaded numbers refer to the imaginary reuse distances (I-RDs) computed from the imaginary text, composed of the unique tokens in reverse order of their first occurrences.

Finally, we extract a **token-level nDMD (T-nDMD)** feature, which collects the nDMD values at the granularity of each token, summarizing each token's recurrence patterns. This fine-grained representation allows us to capture lexical variation in reuse behavior within a document, complementing the global statistics.

3 Experimental Setups

3.1 Datasets and Domains

Most experiments use the English subset of the M4 dataset (Wang et al., 2024), which contains paired human-written and machine-generated texts across five domains (Wikipedia, WikiHow, ArXiv, PeerRead, Reddit) and six language models: ChatGPT (OpenAI, 2022), BLOOMz (Muennighoff et al., 2022), Davinci003 (OpenAI, 2023), Dolly-v2 (Conover et al., 2023), Cohere (Cohere, 2023), and FLAN-T5 (Wei et al., 2022). We focus on English for its analytic structure that preserves locality patterns, and for the dataset's broad model and domain coverage. For MGT classification, we adopt a single-model, single-domain setup with a 7:3 random train–test split. To complement this controlled benchmark, we also evaluate under an open-domain setting using the train and test sets from the GenAIDetect shared task 1a (Wang et al., 2025), which include much more heterogeneous human and model-generated texts, reflecting practical detection scenarios.

3.2 Tokenization and Feature Extraction

We use the pretrained Llama 3 (Grattafiori et al., 2024) BPE tokenizer (Sennrich et al., 2016) as the primary tokenization scheme. To examine the effect of tokenization granularity, we also test the GPT-2 (Radford et al., 2019), word-level SpaCy[2] (Honnibal et al., 2020), and NLTK (Bird and Loper, 2004) tokenizers. After tokenization, we compute reuse distances (RDs) for each document's token stream and derive normalized data movement distance (nDMD) and token-level nDMD (T-nDMD) as features for analysis. We further extract the distributional moments (mean, variance, skewness, kurtosis) of RDs in each document under different formulations to evaluate their class-separation power. Lastly, we collect token frequency counts for each document (bag-of-words) as baselines for comparison. Code related to this work is available at `https://github.com/YYWmus/locality-metrics-for-text`.

For fixed-window nDMD, we randomly sample 1000 substrings for each tested window length from each M4 subset and compute nDMD on these windows. Subsets with insufficient support (< 10 documents with lengths at least $1.5\times$ the window size) are excluded.

3.3 Evaluation Methods

In the characterization, we report the statistical mean and standard deviations of nDMD statistics, and show the significance via t-test p values where appropriate. Classification experiments use logistic regression as the classifier, and report the mean accuracy and/or macro-F1 over 20 seeds. All computations are implemented in Python and executed on CPU servers (Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz), emphasizing the lightweight and efficient nature of the proposed approach.

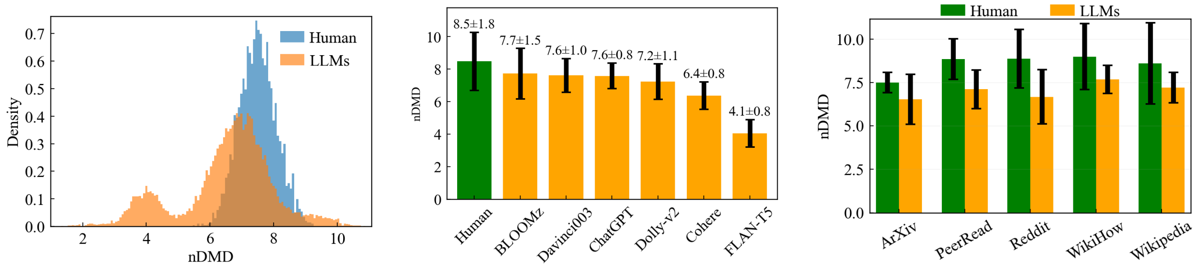
4 Results and Analysis

We evaluate the proposed locality metrics through three complementary analyses: (1) characterization of human and machine text locality patterns, (2) examination of design choices in RD computation, and (3) classification performance using locality features and baselines.

4.1 Characterization

Document-level nDMD. We first characterize the nDMD difference between human- and LLM-generated text computed over entire documents (Figure 2). The raw density plot shows that human text exhibits an approximately normal nDMD distribution with a well-defined peak. In contrast,

[2] Using the `en_core_web_sm` pipeline.



(a) Exemplary nDMD distribution of Human and LLM text in ArXiv domain.

(b) nDMD statistics (mean ± std) for Human and each LLM across domains.

(c) nDMD statistics (mean ± std) for Human and LLMs in each domain.

Figure 2: Document-level nDMD distributions.

LLM-generated text (aggregated across all six models) displays a multi-modal distribution, reflecting systematic differences in LLMs' generation behaviors.

When further grouped by source (Figure 2b) or by domain (Figure 2c), human text consistently attains higher mean nDMD and larger variance than any LLM counterpart, indicating longer-range token reuse and greater lexical diversity. The ArXiv (academic abstract) domain is the sole outlier: human text exhibits substantially lower variance than machine text and also the smallest mean nDMD among all domains. This pattern is likely attributable to the concise, highly standardized nature of abstract-style writing. Statistical tests confirm that, except for a single outlier (ArXiv–BLOOMz), human text has significantly higher nDMD than machine-generated text across all subsets of M4 ($p < 0.001$).

Fixed-window nDMD Since nDMD tends to increase with text length, we control for this effect by performing a fixed-window analysis. Specifically, we randomly sample substrings of fixed lengths from tokenized documents and compute nDMD on these windows. Figure 3 shows the growth of the mean nDMD with window length, with shaded regions indicating the standard deviation. Across domains (Figure 3a), human nDMD increases more rapidly than LLM nDMD, and the separation becomes more pronounced for longer windows.

Notably, some models (e.g., BLOOMz and Cohere) fail to exhibit sustained nDMD growth as the window length increases. This behavior is partly attributable to generation collapse. For example, BLOOMz occasionally produces long sequences of numbers. An instance from PeerRead (ID: 517) reads:

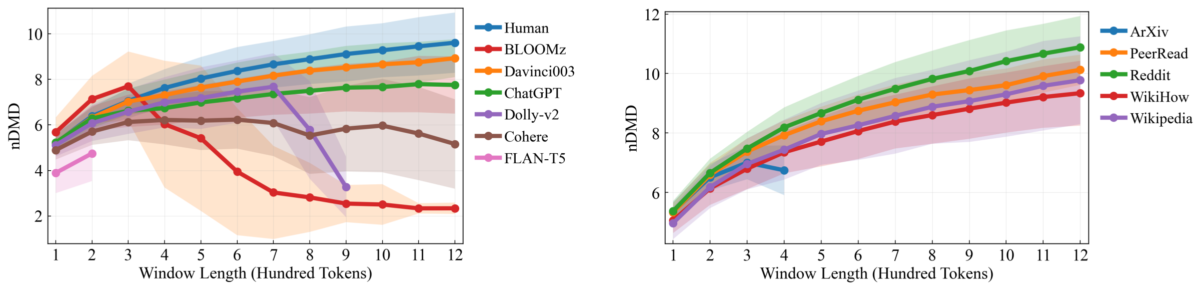
> "…so I'll give 4 stars instead of 3. 4.4.1.2.3.4.4.4.4 1 2 3 4 5 6 …"

The sequence then enumerates until "249", where the generation terminates. Such collapse events drastically reduce nDMD by introducing a large number of extremely short reuse distances from repetitive tokens, thereby lowering the mean and inflating the variance. This discovery highlights nDMD as an effective diagnostic for identifying and analyzing generation collapse in LLMs.

Finally, domain-separated analysis of human text (Figure 3b) shows that nDMD variation reflects writing style rather than document length alone. More formal or task-oriented domains targeting general audiences (e.g., ArXiv, WikiHow, Wikipedia) exhibit lower nDMD, whereas domains characterized by interactive or audience-specific communication (e.g., Reddit and PeerRead) tend to have higher nDMD. The ArXiv domain also features a limited length, which helps explain the observed nDMD outlier in the ArXiv–BLOOMz pair, as BLOOMz exhibits particularly high nDMD for short text while collapsing quickly as the generation extends (Figure 3a).

4.2 Effect of RD Variants

To assess design choices in computing reuse distance (RD), we compare six RD variants formed by combining two distance functions (linear vs. square-root) and three treatments of first occurrences: OMIT (exclude first occurrences), ZERO (assign RD = 0), and I-RD (assign incremental RD values, as illustrated in Figure 1). Using the RD distribution moments (mean, variance, skewness, kurtosis) and the first-use ratio ($\frac{|A|}{|D|+|A|}$) of each variant as features in a logistic regression classifier, we find that the $\sqrt{\text{RD}}$ formulation with I-RD treatment yields the strongest separation between human and machine text (Table 1). One-sided sig-



(a) nDMD vs. window length for Human and each LLM.

(b) Human text nDMD vs. window length for each domain.

Figure 3: Fixed-window nDMD.

nificance tests (H_1: $\sqrt{\text{RD}}$ > linear; I-RD > OMIT; I-RD > ZERO) confirm that this configuration, which is used in our nDMD computation, performs significantly better.

	linear-RD	$\sqrt{\text{RD}}$	I-RD >
OMIT	85.7	87.8	0.02
ZERO	85.3	86.1	<0.001
I-RD	86.2	**88.6**	-
$\sqrt{\text{RD}}$ >	<0.001	-	

Table 1: MGT classification accuracies (mean across domains and LLMs) using six RD variants and p-values of one-sided paired t-tests.

4.3 Machine-Generated Text (MGT) Classification

We compare token-level nDMD (T-nDMD) with a bag-of-words (BoW) baseline and their concatenated combination (Combined). T-nDMD and BoW share the same feature space, one metric per token, making the pair well-suited for testing the robustness of locality statistics against frequency-based ones.

In single-model, single-domain experiments across four tokenizers, Combined features achieve the highest classification accuracy, with fine-grained subword tokenizers (LLaMA3, GPT2) outperforming word-level ones (SpaCy, NLTK) for both locality and frequency features (Table 2). Although the differences are smaller, T-nDMD alone surpasses BoW on mean accuracy in three of four tokenizers, except NLTK. In fact, all methods show a notable drop in performance with NLTK, likely because it treats hyphenated words as single tokens, resulting in the coarsest tokenization granularity.

In open-domain experiments with GenAI Detection Task 1a datasets, overall accuracy drops due to the shift in domains and the diverse MGT collection methods in the test set. However, the combined features still perform the best, and T-nDMD alone surpasses BoW by a larger margin (Table 3).

Features	LLaMA3	GPT-2	SpaCy	NLTK
B: BoW	98.6	98.6	97.9	96.6
T: T-nDMD	98.9	99.0	98.9	95.4
C: Combined	**99.3**	**99.2**	**99.1**	**97.0**
H_1: T > B	0.107	0.052	0.008	0.997
H_1: C > B	<0.001	<0.001	<0.001	0.026

Table 2: Mean MGT classification accuracies and p-values from one-sided significance tests

Features	F1 (macro)	Accuracy
B: BoW	63.1	66.3
T: T-nDMD	65.1	67.8
C: Combined	**66.8**	**69.0**

Table 3: Open-Domain MGT Classificaion macro F1 and Accuracy.

5 Conclusion

In this paper, we introduce locality metrics as a new analytic perspective for studying text, extending ideas from systems locality to natural language. Specifically, we adapt a locality metric for text analysis by defining a normalized version, nDMD, with a new term essential for characterizing low-frequency tokens. Our evaluation of the English M4 and GenAI detection datasets has shown that nDMD values are systematically higher in human text than in machine text. Furthermore, nDMD-based features outperform frequency baselines and enhance performance when they are integrated. Finally, the new normalization treatment is more effective in discriminating between human and machine text than the original DMD and its other extensions.

Limitations

Our analysis is currently limited to English text, where token boundaries and analytic syntax make locality patterns more interpretable. Extending this approach to morphologically rich or synthetic languages may require adapting the definition of reuse and distance to account for complex inflectional structures.

The study also covers a restricted set of domains and language models, many of which are relatively dated. Future evaluations should include newer LLMs and broader text genres to test the robustness of locality statistics under more diverse generative conditions.

While locality features perform on par with or better than frequency-based features in controlled settings, their classification accuracy drops substantially in open-domain scenarios. For comparison, a fine-tuned RoBERTa baseline in the GenAI Detection shared task (Wang et al., 2025) achieved an F1 score of 73.4% on the test set, which is significantly higher than any purely statistical features we tested. Our goal in presenting classification results is to demonstrate that locality statistics differ systematically between human and machine text, and that these differences lead to linearly separable classes, even with stronger discriminative power than frequency-based statistics. A promising direction for future work is integrating locality statistics into neural methods to evaluate whether they enhance existing neural baselines. This work provides foundational experimentation and characterization to motivate such investigations.

Acknowledgement

The first author was supported in part by the CS department, the Hajim school, and a GIDS-AI Seed Fund. Boyang Wang participated in the early stages of the project, contributing preliminary measurements on multiple datasets and developing analytical scripts, including efficient Python implementations for reuse distance (RD) collection. Jack Cashman contributed in project meetings and discussions. In a separate study with the second author (Ding et al., 2023), Toranosuke Ozawa and Wesley Smith discussed possible DMD definitions for first-time data accesses, including the one used by nDMD.

References

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: Can chatgpt detect ai-generated text? *SIGKDD Explor. Newsl.*, 25(2):14–21.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Cohere. 2023. Cohere large language models. `https://cohere.com`. Model documentation available at https://docs.cohere.com/. Accessed: 2026-01-24.

Mike Conover, Ben Kuester, Luke Miller, and 1 others. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`. Databricks technical report.

Chen Ding, Christopher Kanan, Dylan McKellips, Toranosuke Ozawa, Arian Shahmirza, and Wesley Smith. 2023. Dmc4ml: Data movement complexity for machine learning. *Preprint*, arXiv:2312.14441.

Chen Ding and Yutao Zhong. 2003. Predicting whole-program locality with reuse distance analysis. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, San Diego, CA.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21258–21266.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

R. L. Mattson, J. Gecsei, D. Slutz, and I. L. Traiger. 1970. Evaluation techniques for storage hierarchies. *IBM System Journal*, 9(2):78–117.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your large language models are leaving fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. *Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT*, page 152–170. Springer Nature Singapore.

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*, ICML'23. JMLR.org.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Frank Olken. 1981. Efficient methods for calculating the success function of fixed space replacement policies. Technical Report LBL-12370, Lawrence Berkeley Laboratory.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/index/chatgpt/`. Accessed: 2026-01-24.

OpenAI. 2023. Gpt-3.5 model documentation. `https://platform.openai.com/docs/models/gpt-3-5`. Accessed: 2026-01-24.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Wesley Smith, Aidan Goldfarb, and Chen Ding. 2022. Beyond time complexity: data movement complexity analysis for matrix multiplication. In *Proceedings of the International Conference on Supercomputing*, pages 32:1–32:12. ACM.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multilingual 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, St. Julian's, Malta. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, and 7 others. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):1–39. Publisher: BioMed Central.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Yu, Brian Lester, Nan Du, Andrew Dai, and Quoc Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Jake Wires, Stephen Ingram, Zachary Drudi, Nicholas JA Harvey, Andrew Warfield, and Coho Data. 2014. Characterizing storage workloads with counter stacks. In *Proceedings of the Symposium on Operating Systems Design and Implementation*, pages 335–349. USENIX Association.

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.

Liang Yuan, Chen Ding, Wesley Smith, Peter J. Denning, and Yunquan Zhang. 2019. A relational theory of locality. *ACM Transactions on Architecture and Code Optimization*, 16(3):33:1–33:26.

Yutao Zhong, Xipeng Shen, and Chen Ding. 2009. Program locality analysis using reuse distance. *ACM Transactions on Programming Languages and Systems*, 31(6):20:1–20:39.

A Document-level nDMD statistics

In this appendix, we display the aggregated nDMD statistics (mean ± std) of human and machine-generated text for all four tokenizers used in the experiment (Figure 4).

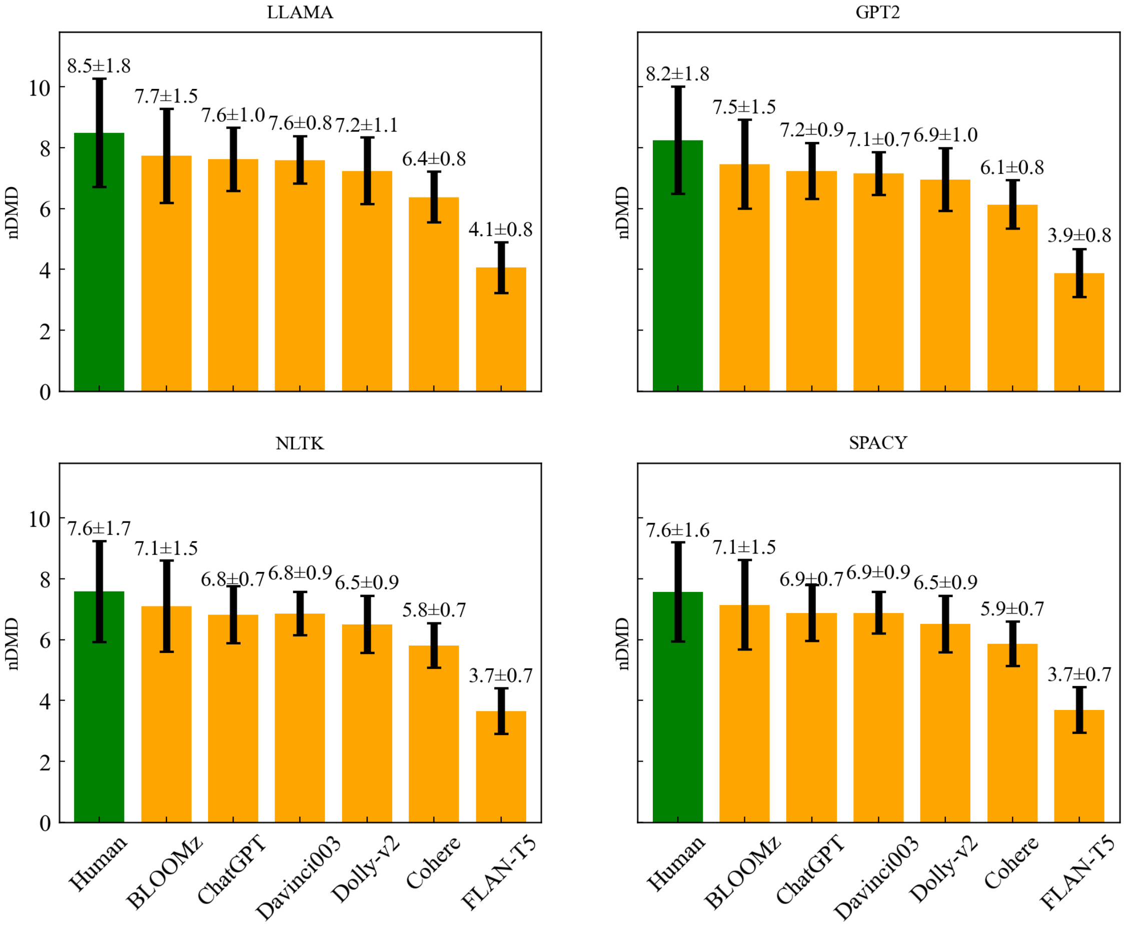


Figure 4: Mean and standard deviation of nDMD across domains.

The results from these tokenizers further support the previous finding that human text exhibits higher nDMD and greater variance than LLM-generated text. We also observe that the gap between human and the closest matched LLM (BLOOMz) is larger with subword tokenizers (Llama 3 and GPT-2), which may help explain the subword tokenizer's superior performance in the MGT classification using T-nDMD features (Table 2).

In the tables below (Table 4-7), we further list the nDMD statistics separated by model and domain. Unless otherwise noted, all pairwise differences are statistically significant ($p < 0.05$ in one-sided independent t-test H_1 : Human $>$ LLM). A superscript † indicates $p \geq 0.05$. To summarize, Arxiv-BLOOMz is a consistent outlier, and a few other domain-model pairs using the SpaCy tokenizer (in ArXiv and/or using BLOOMz) failed the significance tests. The rest of the experiments show patterns consistent with aggregated statistics.

Table 4: Mean ± standard deviation of nDMD separated by model and domain, with **Llama 3** tokenizer.

	ArXiv	PeerRead	Reddit	WikiHow	Wikipedia	All domains
BLOOMz	**8.19 ± 0.96**†	8.26 ± 0.85	8.30 ± 1.16	8.31 ± 1.18	5.98 ± 1.49	7.72 ± 1.55
ChatGPT	7.14 ± 0.43	7.80 ± 0.43	7.28 ± 0.70	8.14 ± 0.95	7.74 ± 0.55	7.59 ± 0.78
Cohere	6.14 ± 0.59	6.67 ± 0.64	5.91 ± 1.04	6.52 ± 0.84	6.66 ± 0.83	6.37 ± 0.84
Davinci003	6.70 ± 0.59	6.32 ± 0.49	7.57 ± 0.76	8.30 ± 1.09	8.15 ± 0.64	7.62 ± 1.03
Dolly-v2	7.10 ± 0.79	8.09 ± 1.13	7.08 ± 0.98	7.15 ± 1.13	7.48 ± 1.29	7.24 ± 1.09
FLAN-T5	3.94 ± 0.62	5.47 ± 1.51	3.90 ± 0.52	–	–	4.05 ± 0.84
All LLMs	6.53 ± 1.44	7.10 ± 1.12	6.67 ± 1.57	7.68 ± 0.81	7.20 ± 0.87	6.76 ± 1.42
Human	7.49 ± 0.58	**8.85 ± 1.17**	**8.86 ± 1.69**	**8.98 ± 1.90**	**8.60 ± 2.34**	**8.47 ± 1.78**

Table 5: Mean ± standard deviation of nDMD separated by model and domain, with **GPT-2** tokenizer.

	ArXiv	PeerRead	Reddit	WikiHow	Wikipedia	All domains
BLOOMz	**8.11 ± 0.93**†	8.02 ± 0.92	7.97 ± 1.13	7.75 ± 1.10	5.86 ± 1.41	7.45 ± 1.46
ChatGPT	6.75 ± 0.39	7.40 ± 0.41	6.79 ± 0.53	7.76 ± 0.83	7.23 ± 0.45	7.15 ± 0.70
Cohere	5.94 ± 0.56	6.41 ± 0.61	5.56 ± 0.96	6.20 ± 0.76	6.50 ± 0.79	6.13 ± 0.79
Davinci003	6.45 ± 0.58	6.07 ± 0.46	7.10 ± 0.62	7.87 ± 0.97	7.71 ± 0.55	7.22 ± 0.92
Dolly-v2	6.90 ± 0.75	7.89 ± 1.07	6.70 ± 0.87	6.77 ± 1.07	7.26 ± 1.24	6.95 ± 1.04
FLAN-T5	3.79 ± 0.57	5.28 ± 1.44	3.69 ± 0.46	–	–	3.88 ± 0.79
All LLMs	6.32 ± 1.43	6.85 ± 1.10	6.30 ± 1.50	7.27 ± 0.75	6.91 ± 0.73	6.46 ± 1.35
Human	7.29 ± 0.57	**8.60 ± 1.21**	**8.49 ± 1.63**	**8.68 ± 1.78**	**8.56 ± 2.42**	**8.24 ± 1.75**

Table 6: Mean ± standard deviation of nDMD separated by model and domain, with **SpaCy** tokenizer.

	ArXiv	PeerRead	Reddit	WikiHow	Wikipedia	All domains
BLOOMz	**7.51 ± 0.85**†	7.94 ± 0.82†	7.69 ± 1.10	7.80 ± 1.01	5.40 ± 1.42	7.14 ± 1.48
ChatGPT	6.41 ± 0.34†	7.09 ± 0.36	6.65 ± 0.49	7.59 ± 0.80	6.82 ± 0.38	6.88 ± 0.68
Cohere	5.58 ± 0.50	6.14 ± 0.53	5.39 ± 0.88	6.13 ± 0.73	6.04 ± 0.69	5.86 ± 0.73
Davinci003	5.91 ± 0.50	5.87 ± 0.39	6.90 ± 0.56	7.64 ± 0.94	7.22 ± 0.46	6.87 ± 0.91
Dolly-v2	6.36 ± 0.66†	7.37 ± 0.96	6.44 ± 0.80	6.41 ± 0.99	6.66 ± 1.12	6.51 ± 0.93
FLAN-T5	3.52 ± 0.53	5.06 ± 1.30	3.58 ± 0.43	–	–	3.69 ± 0.74
All LLMs	5.88 ± 1.33	6.58 ± 1.07	6.11 ± 1.44	7.11 ± 0.78	6.43 ± 0.71	6.16 ± 1.29
Human	6.36 ± 0.47	**7.98 ± 0.96**	**7.94 ± 1.36**	**8.27 ± 1.64**	**7.75 ± 2.15**	**7.56 ± 1.63**

Table 7: Mean ± standard deviation of nDMD separated by model and domain, with **NLTK** tokenizer.

	ArXiv	PeerRead	Reddit	WikiHow	Wikipedia	All domains
BLOOMz	**7.42 ± 0.88**†	7.87 ± 0.87	7.71 ± 1.12	7.75 ± 1.03	5.34 ± 1.41	7.09 ± 1.49
ChatGPT	6.32 ± 0.33	7.06 ± 0.37	6.60 ± 0.48	7.63 ± 0.82	6.80 ± 0.38	6.85 ± 0.72
Cohere	5.49 ± 0.50	6.07 ± 0.54	5.35 ± 0.88	6.09 ± 0.73	6.01 ± 0.69	5.80 ± 0.74
Davinci003	5.82 ± 0.49	5.77 ± 0.40	6.85 ± 0.56	7.62 ± 0.95	7.18 ± 0.46	6.81 ± 0.94
Dolly-v2	6.30 ± 0.66	7.39 ± 0.96	6.44 ± 0.81	6.42 ± 0.99	6.65 ± 1.13	6.49 ± 0.94
FLAN-T5	3.46 ± 0.53	5.00 ± 1.31	3.58 ± 0.43	–	–	3.65 ± 0.74
All LLMs	5.80 ± 1.32	6.52 ± 1.09	6.09 ± 1.45	7.10 ± 0.78	6.39 ± 0.73	6.12 ± 1.29
Human	6.39 ± 0.52	**8.03 ± 1.02**	**7.92 ± 1.38**	**8.27 ± 1.68**	**7.80 ± 2.20**	**7.58 ± 1.66**

B Document-level nDMD distributions

We present the raw nDMD distribution histograms for each domain as well as for all domains combined (Figure 5 and 6). Human text generally follows an approximately normal distribution, with occasional right-skewed tails arising from imbalanced document lengths in certain domains. As shown (Figure 6), individual LLMs also exhibit quasi-normal distributions within each domain–model subset. When aggregated across models, however, LLM-generated text forms multi-modal distributions, reflecting the tendency of different models to concentrate in distinct nDMD ranges.

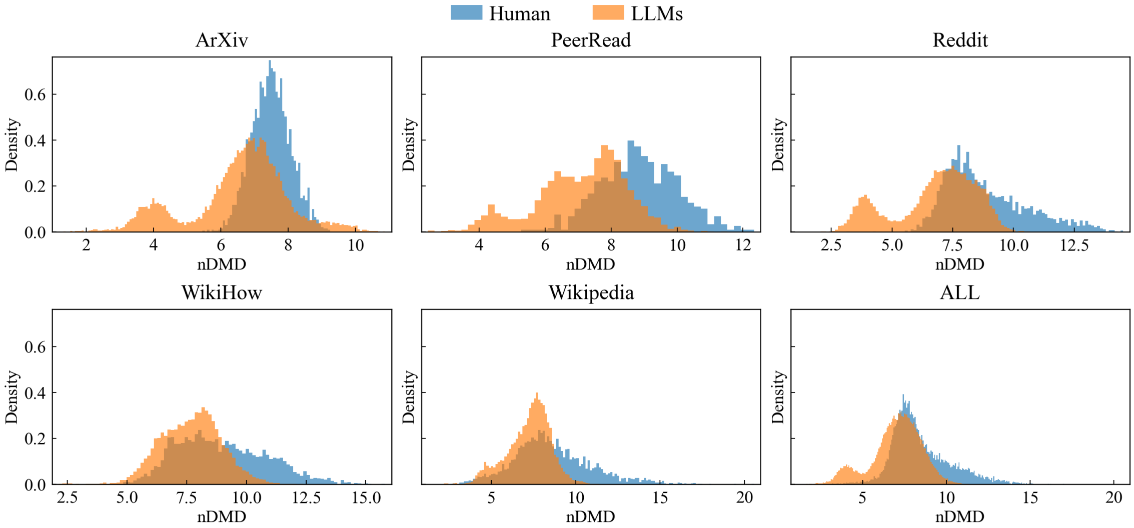


Figure 5: nDMD distributions by domain, Human and LLMs (aggregated).

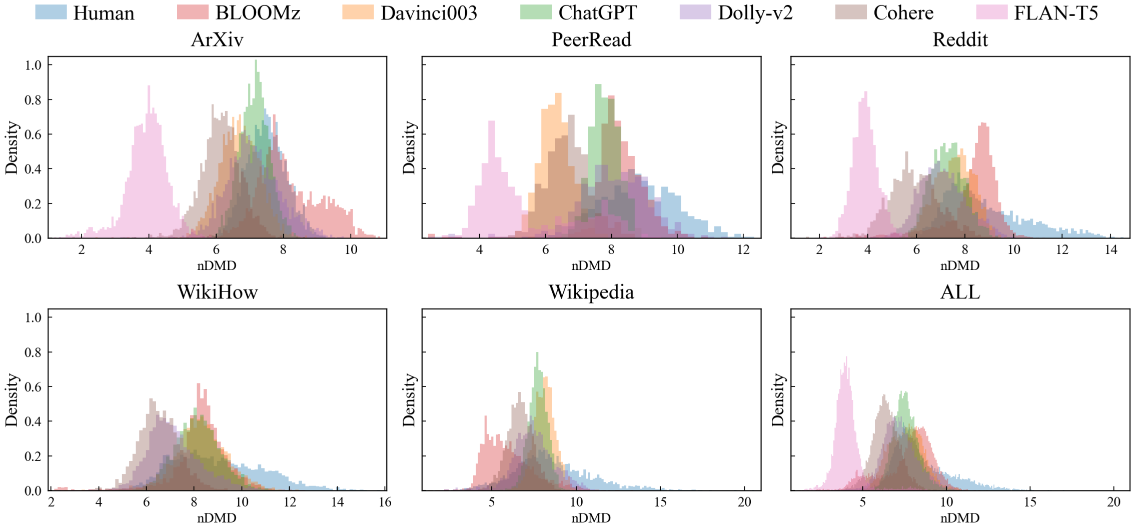


Figure 6: nDMD distributions by domain, Human and each LLM.

C Fixed-window nDMD in each domain

We further present fixed-window nDMD scaling for each domain (Figure 7). The aggregated plots reveal text-length limitations in certain domains and models; for LLMs, shorter effective lengths are partly attributable to certain generation prompts that explicitly constrain output length. Notably, ArXiv contains particularly short texts in both human- and model-generated data.

Across domains, human text consistently exhibits higher nDMD than LLM-generated text at larger window sizes, with ArXiv as the sole exception. In this domain, human text also shows an atypical downturn at the longest window length, likely reflecting the highly standardized and tightly constrained nature of abstract-style writing with highly concentrated content.

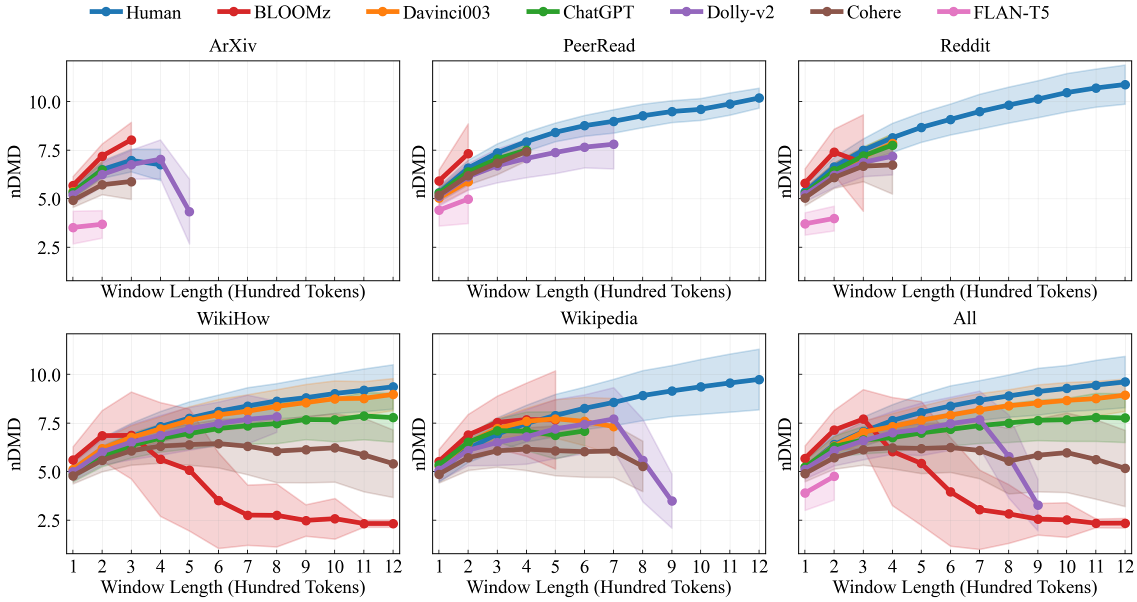


Figure 7: Fixed-window nDMD trends, all domains.