Robust Statistical Detection of Short and Paraphrased AI-Generated Text

Tanishq Choudhary
Department of Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
tanishqchoudhary353@gmail.com

***Abstract*—The ever-increasing rise of AI-generated short texts across the internet has made their reliable detection an urgent priority. Identifying such content has become especially difficult in situations when the text length is limited to 50–200 tokens, a range in which many existing detection methods fall short due to a reliance on longer inputs. It becomes harder when AI generated texts are paraphrased or rewritten because such transformations can severely diminish the effectiveness of current detectors. This paper explores the robustness of statistical and stylometric indicators of detection against short and paraphrased AI-generated text without relying on large neural classifiers. In this work, we create a dataset consisting of human-written samples, outputs from multiple modern language model families, and paraphrased variants of each-all controlled within the same length range. Then, using lexical diversity measures, function-word patterns, part-of-speech distributions, psycholinguistic cues, and other lightweight statistical features, we assess how well these signals can survive paraphrasing. Our results show improved robustness compared to perplexity-based baselines, consistent performance across model families, and practicality on modest hardware. The findings point to a viable path toward reliable detection of short and rewritten AI-generated content.**

***Index Terms*—AI-generated text detection, Stylometry, Short text detection, Paraphrasing robustness, Statistical NLP**

I. Introduction

Large language models have been increasingly employed to produce snippets of text indistinguishable from human writing. These models are now contributing to content across social media platforms, academic settings, customer support systems, and general online communication. With the surge in their adoption, differentiating between text produced by humans and that generated by machines has emerged as a pressing issue in applications requiring authorship verification, content authenticity, and integrity monitoring [1]–[3].

While there has been significant progress in developing automated detectors, most of the existing work focuses on longer passages, where statistical irregularities are easier to catch. Most detection performances decrease drastically when text length falls below 200 tokens, and many detectors implicitly rely on long-form structure or context to achieve meaningful accuracy. Therefore, short segments, typical of posts, comments, answers, and micro-communications, are a very challenging setting, not well explored as yet [4]–[6].

The challenge becomes more severe when text is paraphrased or rewritten. Even mild paraphrasing can obscure surface patterns that different approaches to detecting generators, such as perplexity-based and classifier-based detectors, rely on. Indeed, recent findings have shown that rewriting, style transfer, and grammar polishing can consistently degrade detector accuracy, sometimes rendering them ineffective. However, a systematic study of detection robustness under paraphrasing, especially within short lengths of text, remains limited [7]–[9].

Most of the existing detection approaches rely on large neural classifiers or model-specific cues. These systems may generalize poorly across model families and can require substantial computation. Watermarking-based solutions provide complementary benefits but are usually missing in real-world scenarios, might be easily removed by rewriting, and are not yet implemented consistently across the models [10]–[13].

This work examines an alternative: focusing on statistical, stylometric, and psycholinguistic features that are far more stable and interpretable, even under paraphrasing. Our study focuses specifically on the 50–200 token range of texts, where most detectors fail and where practical applications abound. We take a feature-driven approach, forgoing the need to train large neural networks to analyze signals such as lexical diversity, function-word patterns, part-of-speech distributions, readability indicators, and other lightweight metrics calculable on modest hardware.

To establish this assessment, we create a controlled dataset consisting of human-written samples, outputs from various LLM families, and various paraphrased versions of each. This allows for systematic testing across models, paraphrasing intensities, and content domains. Our analysis indicates which features remain reliable under rewriting, how they behave across model families, and how they compare to perplexity-based baselines. Taken together, this work presents a clear indication of the stylometric and statistical cues essential for the robust detection of short-text AI. These findings give insight into real-world detection scenarios involving paraphrased or edited content and form the basis of developing methods that remain effective as LLM-generated text becomes increasingly common.

II. Related Work

A. AI-Generated Text Detection

Research in the area of AI-generated text detection has grown rapidly along with large language models' increased capability and widened deployment. Wu et al. present DetectRL,

a benchmark for evaluating detection robustness under real-world conditions such as prompt manipulations, human edits, and diverse writing styles [14]. Their results are that state-of-the-art detectors still fail in practical scenarios, especially when adversarial or noisy modifications are introduced. Hu et al. put forward RADAR, a framework trained adversarially against a paraphraser to improve robustness to rewritten content across various LLM families [15]. Akram empirically evaluated commercially used detectors and found significant variations in performance between domains and sometimes even well-known tools can misclassify AI-generated text with unexpected frequency [16]. All these studies together point out the limitations of detectors, more so when applied to diverse domains and under adversarial conditions.

B. Stylometry and Authorship Attribution

Stylometric methods provide yet another direction in understanding the writing patterns that distinguish human authors from LLMs. Classic work by Calle-Martin and Miranda-Garcia outlines the foundations of stylometry and its applications for authorship attribution [17]. Recently, this thread of work gained new relevance due to modern neural text generators. Kumarage and Liu explored neural authorship attribution for distinguishing proprietary versus open-source LLMs, showing that both stylometric and linguistic features can complement neural classifiers and provide interpretable signals [18]. Bisztray et al. extended the stylometric analysis to the code domain by developing the LLM-AuthorBench dataset and demonstrating that code generated by LLMs exhibits model-specific stylistic fingerprints [19]. Most recently, Huang et al. presented an extensive survey of authorship attribution in the era of LLMs by categorizing human, machine, and hybrid authorship scenarios; they stressed that there is an increasing need for methods that generalize well across domains and model families [20]. These works highlight the relevance of stylometric and psycholinguistic features as lightweight, interpretable means of distinguishing between human and machine-generated content.

C. Paraphrasing, Robustness, and Adversarial Attacks

One of the most enduring issues with the detection of AI-text has to do with robustness to paraphrasing. Krishna et al. showed that, even when the actual semantics of the text are unchanged, paraphrasing can reliably evade many detection systems [7]. Cheng et al. introduced a universal adversarial paraphrasing attack which rewrites machine-generated text to appear more humanlike while explicitly targeting detector weaknesses [8]. Shportko and Verbitsky assessed a range of detection methods for resistance to paraphrasing attacks and uncovered substantial variation in robustness across model types and detection strategies [9]. Such findings point out a crucial need for feature sets that preserve discriminative signals under rewriting—a gap that motivates the stylometric and statistical focus of this paper.

D. Model-Specific Detection and LLM Attribution

Going beyond binary classification, namely, human vs. machine-generated text, various works have tried to identify which LLM generated a given text sample. Yildiz et al. benchmarked LLMs and LLM-based agents on practical vulnerability detection in code repositories, showing that different models leave identifiable behavioral signatures when reasoning over code [21]. Though their work pertains specifically to software vulnerabilities, it underlines an important aspect: model families often produce recognizable patterns. Bisztray et al. further showed that even highly capable LLMs leave stylistic traces in generated code, which can be used for model attribution [19]. Huang et al.'s survey places such tasks within a broader taxonomy of authorship attribution challenges in the LLM era [20]. Collectively, this research suggests that robust statistical and stylistic cues may give signals that are valid across different types of content.

III. 3. Methodology and Framework

This section describes the general framework to detect short AI-generated and paraphrased text. The approach relies on lightweight statistical, stylometric, and psycholinguistic features instead of using large neural classifiers. All analysis is done on a fixed set of topic-aligned text samples containing human-written Wikipedia passages and the LLM-generated responses for the same topics.

A. System Architecture

The system is designed as a modular pipeline that processes human and AI-generated text samples in several stages.

Data Collection:

A range of topics was chosen from Wikipedia, including, among others, the Theory of Relativity, Modernist Literature, the Marvel Cinematic Universe, the Fall of the Roman Empire, Existentialism, CRISPR gene editing, and Blockchain Technology. For each of these topics, a human-written reference passage was saved as a .txt file; for the same topics, prompts were fed into three different LLMs: DeepSeek, ChatGPT, and Bard. Their respective outputs were also saved in the dataset folder.

Preprocessing:

All text files were normalized by removing formatting artifacts, unnecessary whitespace, HTML remnants (if any), and other noise. Token length was restricted to the 50–200 token range. There were no semantic changes other than what is done by the models used.

Feature Extraction:

For each text sample, stylometric and statistical features were calculated: measures of lexical diversity, ratios of function words, distribution of parts-of-speech, readability indicators, character-level entropy, and compression-based features. The aforementioned set forms the fixed-length representation of a document.

Classification Module:

These feature vectors were passed to lightweight classifiers comprising logistic regression and random forest models. No

finetuning or large-scale training was needed; the classifiers operate purely on the precomputed features.

Evaluation:

Performance in the system is evaluated along various dimensions: human vs. AI, cross-model generalization, and robustness to paraphrasing. Detection quality is measured as accuracy, precision, recall, F1-score, ROC-AUC, and degradation under paraphrased conditions.

B. Algorithms and Techniques

The methodology is based on algorithms which are inexpensive in computation and interpretable.

Lexical Diversity Metrics:

Type-token ratio variants, such as MSTTR, Herdan's C, and Honoré's statistic are measures that quantify vocabulary variation in short texts.

Function-word statistics:

The distribution of high-frequency grammatical words is analyzed, since these patterns tend to be stable under paraphrasing.

Part-of-Speech Patterns:

POS tag ratios and bigrams help capture the sentence structure irrespective of the words that are used.

Readability and Psycholinguistic Indicators:

Metrics such as Flesch-based readability and concreteness estimates show the stylistic properties of the text

Entropy and Compression Features: Character-level entropy and normalized compression size are indicators of repetitiveness and predictability which are features that generally differ in human and AI-generated writing. Baseline Perplexity: A small LM is used only for comparisons based on perplexity and is not considered a main detection method.

C. eDetailed Design Methodologies

All components were arranged in a pipeline to ensure that all results are reproducible.

Topic-Relevant Samples: Using the same set of topics across human-written text and model-generated responses reduces content variance and isolates stylistic differences.

Fixed-Length Processing: All text samples have been shortened/picked to stay within the 50-200 word token range.

Lightweight Model Design: Only traditional machine-learning algorithms that have low computational needs are used, hence no powerful gpus are required and conventional commerical hardware can be used.

Cross-Model Evaluation: This design has comparisons across the outputs from DeepSeek, ChatGPT and Bard to ensure cross model compatibility.

Paraphrase Robustness Testing: Explicitly, paraphrased outputs are compared against the same feature sets to see which signals remain stable under rewriting.

IV. Experimental Analysis Summary

This pipeline is designed for the analysis of whether short AI-generated text can be differentiated from human-written text by transparent, interpretable linguistic measures rather than large machine-learning models. In essence, the aim is to find out through quantitative stylometry and statistical testing whether systematic differences exist between Wikipedia passages and the outputs of three LLM families, namely DeepSeek, ChatGPT, and Bard.

The feature extraction process focuses on five groups of linguistic signals. The first is information-theoretic behavior, where entropy and compression metrics capture how predictable a text is. Because language models tend to produce high-probability sequences, their outputs are expected to compress more efficiently than human-authored writing. The second group focuses on structural burstiness, measured as the variation in sentence lengths. Human text usually shows large fluctuations, while the generated sentences of LLMs tend to be of more uniform length. The third group measures lexical richness through type–token statistics and counts of words that appear only once. This probes vocabulary diversity across sources. Readability metrics, like the Flesch Reading Ease score, form the fourth group and reflect differences in clarity and density of prose. The fifth group examines basic syntactic morphology through part-of-speech ratios, which allows for a comparison of how often each source uses nouns, verbs, adjectives, and other such categories.

The program uses non-parametric statistical tests to determine if these features show any meaningful separation between classes. A p-value is calculated using the Mann–Whitney U test to see if the differences between AI-generated and human-written text are statistically significant. The effect size is measured using Cohen's d to show how strong the separation is. This integration of significance testing with effect-size measurement provides evidence for which of the stylistic cues remain reliable at short text lengths.

It generates three major artifacts: a CSV file containing the complete feature set for all documents; a table of p-values and effect sizes for all features; and a set of boxplots displaying the feature distributions for each source category. These three artifacts serve as the backbone for the evaluation and discussion sections of the paper.

A. Feature Groups Used in the Study

Table I provides an overview of the feature groups and metrics computed in this study.

TABLE I: Summary of Feature Groups and Metrics Used

Feature Group	Metrics
Information Theory	Entropy, Zlib Compression Ratio
Structural Burstiness	Sentence Length Std. Dev.
Lexical Richness	TTR, Hapax Legomena
Readability	Flesch Reading Ease
Syntactic Morphology	POS Tag Ratios

V. Results and Findings

The current section presents the descriptive patterns observed in the extracted features, followed by the statistical significance tests provided by the Mann–Whitney U procedure and measures of effect size. It seeks to establish whether any

of the computed stylistic indicators will present consistent separation between human-written Wikipedia passages and text generated by DeepSeek, ChatGPT, and Bard.

A. Descriptive Patterns in the Extracted Features

Across all seven topics, there is a general trend in comparing the numeric features: human-written Wikipedia samples generally have noticeably higher values for structural burstiness, reflecting greater variability in sentence lengths. In nearly every topic, Wikipedia's burstiness score is the highest within the group, many times showing values between 10 and 16, whereas the LLM outputs cluster more narrowly between approximately 5 and 9. This pattern is very consistent across all categories, which again means that sentence length variation is a strong stylistic signal.

Other metrics exhibit either weaker or inconsistent separation. Type–token ratio varies for both the human and AI classes, and in several LLM outputs it is higher than Wikipedia, such as the ChatGPT outputs on blockchain and CRISPR. There is substantial overlap of classes in average word length and Shannon entropy; Wikipedia often lies in the middle of the distribution rather than at an extreme. Readability values span a wide range for all sources, demonstrating that the clarity or density of prose depends more greatly on topic than on author.

These descriptive observations suggest that structural features have strong human–AI contrast, whereas lexical and entropy-based features present limited discriminative behavior in short text samples.

B. Statistical Significance of Feature Differences

To quantify these trends, omnibus Mann–Whitney U tests were performed for each feature, considering all human-written samples as one group and all LLM samples as the other group. Effect sizes were computed using Cohen's d.

Among all the metrics evaluated, only structural burstiness has a statistically significant difference, with $p \approx 5.07 \times 10^{-5}$ and a very large effect size ($d \approx 2.33$). This confirms that variation in sentence lengths is indeed the most reliable stylistic cue for distinguishing human and AI writing within the evaluated token limits.

All other characteristics yield non-significant p-values. Sentence length, readability, type–token ratio, entropy, and compression ratio all show largely overlapping distributions between human and machine-generated text. While some of the effect sizes are more moderate—for example, $d \approx 0.57$ for readability—the lack of statistical significance suggests that such differences are not stable enough, or consistent enough across topics, to be reliable indicators.

C. Topic-Level Trends

Looking at raw data over topics, one can see that Wikipedia samples are consistently at the high end of the burstiness range. For example:

- Blockchain Technology: human burstiness = 12.56 vs. AI $\approx$ 7.0–7.7

TABLE II: Mann–Whitney U Significance Tests and Effect Sizes

Feature	P-Value	Significant?	Cohen's d
struct_burstiness	5.07E-05	TRUE	2.33
struct_avg_sent_len	0.155	FALSE	0.61
read_flesch	0.272	FALSE	0.57
lex_avg_word_len	0.348	FALSE	-0.28
lex_ttr	0.604	FALSE	0.12
ent_shannon	0.640	FALSE	0.29
ent_zlib_ratio	0.836	FALSE	0.12

- Fall of the Roman Empire: human burstiness = 16.15 vs. AI $\approx$ 6.8–11.0
- Theory of Relativity: human burstiness = 15.98 vs. AI $\approx$ 6.36–10.47

In contrast, type–token ratio does not show consistent ordering: for several topics, ChatGPT or DeepSeek is higher than Wikipedia, whereas Bard is sometimes considerably lower, such as for Existentialism. Entropy and compression ratio show similar inconsistencies.

These topic-level patterns support the statistical result that burstiness is the only consistently differentiating feature across all scenarios.

The complete feature table containing all extracted metrics for every document is provided in Appendix VII for reference.

D. Visualization of Feature Distributions

The resulting boxplots give a graphical overview of the variation in each feature across the four sources. The burstiness plot exhibits clear separation: the Wikipedia distribution lies substantially above those of all LLMs. By contrast, the boxplots for type–token ratio, average word length, readability, entropy, and compression overlap heavily, confirming the quantitative results.

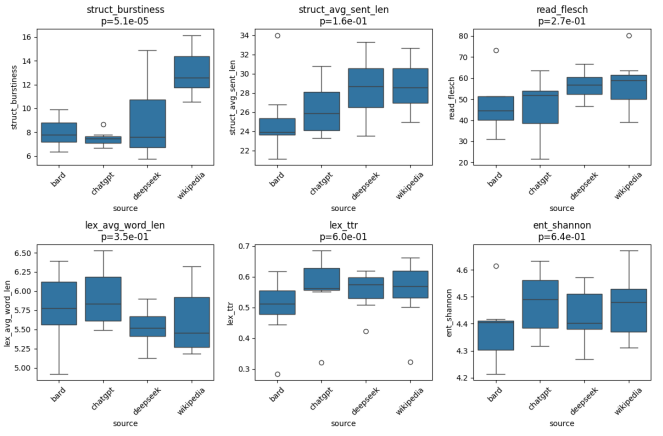

Fig. 1: Distribution of selected features across Bard, ChatGPT, DeepSeek, and Wikipedia.

E. Highlights of the Main Points

According to both descriptive and statistical evaluation, the most reliable discriminator of human vs. AI writing in short texts is structural burstiness. None of the other features based

on lexical, readability, or entropy-based measurements has shown statistically significant separation across all categories. This indicates that short text severely constrains the applicability of traditional stylometric cues and that structural variation is the most robust indicator among the metrics evaluated.

VI. Discussion

These findings emphasize how the performance of the stylometric indicators strongly depends on the length of the text being analyzed. Even though many previous detection methods operate with a lexical richness or entropy-based measure, the present results show that such signals are unstable in short segments with 50–200 tokens. The absence of statistical significance concerning type–token ratio, average word length, readability, and entropy indicates that the variation of these indicators is more related to topic and phrasing than to authorship. Therefore, they cannot be used as reliable ground for short-text detection.

Nevertheless, structural burstiness is a feature that really works across the board. The sentence lengths of human-written passages differ much more from one another, whereas those of all three LLMs result in the same structure regardless of the topic. The large effect size and strong statistical significance suggest that this stylistic trait is highly resilient and generalizes well across model families. This agrees with previous observations that, in optimizing for fluency and coherence, the resulting sentence structures in language models often come out evenly paced.

The limited separability of the remaining features further underlines some of the limitations of short-text scenarios. Short passages reduce vocabulary diversity, narrow the ranges of entropy, and squeeze stylistic variation. Therefore, those features which perform well for long or multi-paragraph content do not work well when applied to shorter text. This suggests that future detection approaches focusing on brief messages, such as social-media posts, chat responses, or comment threads, should rely mainly on structural and rhythm-based features rather than lexical statistics.

Another finding to emerge from the output is the cross-model consistency of the structural patterns. DeepSeek, ChatGPT, and Bard represent very different architectures and scales of training, and yet their outputs show similar sentence-length uniformity, implying that the burstiness-based cues may provide at least some degree of model-agnostic stability. This is an important point, given the rapid turnover of model versions and model families and the fact that detection strategies anchored too tightly within any one LLM tend to grow obsolete.

At the same time, the reliance on this small set of features constitutes a limitation: this approach is intentionally lightweight and does not embed any deeper semantic or contextual cues. Therefore, it is not designed to compete with large neural detectors in terms of accuracy, but more to provide an interpretable and computation-efficient alternative that is customized for short text and paraphrased scenarios. These findings confirm that such a minimalistic setup is still able to isolate at least one strong discriminative signal, even under tight length constraints. Collectively, results reflect both the opportunities and boundaries of feature-based short-text detection. Burstiness provides a robust foundation, but further improvement in performance will likely depend on the combination of structural cues with additional indicators robust to topic, paraphrasing, and text length.

VII. Conclusion

The study shows that the detection of AI-generated text in short segments requires one to move away from the traditional lexical or entropy-based indicators. Among all the evaluated features, only structural burstiness consistently distinguishes human-written passages from the outputs of DeepSeek, ChatGPT, and Bard. The stability of this cue across topics and model families suggests that sentence-level structural variation remains a reliable stylistic marker even when text length is restricted and paraphrasing is present.

At the same time, these results also point to some limitations, which come with relying on lightweight features only. Short text compresses stylistic diversity and diminishes the informative value of many established measures of style, reducing the range of signals one can use. While the approach adopted here intentionally avoids heavy classifiers in favor of interpretability and low computational cost, findings obtained in this study suggest that meaningful detection for short texts should be focused on structural and rhythm-based characteristics, possibly combined with further features that remain stable under topic shifts and rewriting.

The analysis shows that even minimalist stylometric methods can find strong discriminative signals in short text, provided that the features are chosen for robustness rather than tradition. Future work may investigate complementary structural or discourse-level indicators to extend these results and build more resilient detection frameworks.

References

[1] Z. Sun *et al.*, "Are we in the ai-generated text world already? quantifying and monitoring aigt on social media," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[2] A. Gray, "Chatgpt contamination: estimating the prevalence of llms in the scholarly literature," *arXiv preprint arXiv:2403.16887*, 2024.

[3] J. Bevendorff *et al.*, "The two paradigms of llm detection: Authorship attribution vs authorship verification," in *Findings of the ACL 2025*, 2025.

[4] V. S. Sadasivan *et al.*, "Can ai-generated text be reliably detected?" *arXiv preprint arXiv:2303.11156*, 2023.

[5] S. Chakraborty *et al.*, "Position: On the possibilities of ai-generated text detection," in *International Conference on Machine Learning*, 2024.

[6] H. D. S. Gameiro, A. Kucharavy, and L. Dolamic, "Llm detectors still fall short of real world: Case of llm-generated short news-like posts," *arXiv preprint arXiv:2409.03291*, 2024.

[7] K. Krishna *et al.*, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," in *Advances in Neural Information Processing Systems*, 2023.

[8] Y. Cheng *et al.*, "Adversarial paraphrasing: A universal attack for humanizing ai-generated text," *arXiv preprint arXiv:2506.07001*, 2025.

[9] A. Shportko and I. Verbitsky, "Paraphrasing attack resilience of various machine-generated text detection methods," in *Proceedings of the 2025 NAACL Student Research Workshop*, 2025.

[10] Q. Pang *et al.*, "No free lunch in llm watermarking: Trade-offs in watermarking design choices," in *Advances in Neural Information Processing Systems*, 2024.

[11] Y. Liang *et al.*, "Watermarking techniques for large language models: A survey," *arXiv preprint arXiv:2409.00089*, 2024.
[12] J. Liang *et al.*, "Watermark under fire: A robustness evaluation of llm watermarking," *arXiv preprint arXiv:2411.13425*, 2024.
[13] J. Ren *et al.*, "A robust semantics-based watermark for large language model against paraphrasing," in *Findings of the Association for Computational Linguistics: NAACL*, 2024.
[14] J. Wu *et al.*, "Detectrl: Benchmarking llm-generated text detection in real-world scenarios," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 100 369–100 401.
[15] X. Hu, P.-Y. Chen, and T.-Y. Ho, "Radar: Robust ai-text detection via adversarial learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 15 077–15 095.
[16] A. Akram, "An empirical study of ai generated text detection tools," *arXiv preprint arXiv:2310.01423*, 2023.
[17] J. Calle-Martin and A. Miranda-Garcia, "Stylometry and authorship attribution: introduction to the special issue," *English Studies*, vol. 93, no. 3, pp. 251–258, 2012.
[18] T. Kumarage and H. Liu, "Neural authorship attribution: Stylometric analysis on large language models," in *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2023, pp. 1–8.
[19] T. Bisztray *et al.*, "I know which llm wrote your code last summer: Llm generated code stylometry for authorship attribution," *arXiv preprint arXiv:2506.17323*, 2025.
[20] B. Huang, C. Chen, and K. Shu, "Authorship attribution in the era of llms: Problems, methodologies, and challenges," *ACM SIGKDD Explorations Newsletter*, vol. 26, no. 2, pp. 21–43, 2025.
[21] A. Yildiz *et al.*, "Benchmarking llms and llm-based agents in practical vulnerability detection for code repositories," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

APPENDIX: FULL EXTRACTED FEATURE TABLE

TABLE III: Full dataset: extracted feature values for every file (one row per file).

Filename	Source	is_ai	lex_ttr	lex_avg_word_len	read_flesch	struct_avg_sent_len	struct_burstiness	ent_shannon
blockchain_technology_bard.txt	bard	1	0.5589	5.419	44.25	23.91	7.79	4.321
blockchain_technology_chatgpt.txt	chatgpt	1	0.6850	6.528	21.74	24.67	7.45	4.390
blockchain_technology_deepseek.txt	deepseek	1	0.6186	5.549	46.68	24.90	7.03	4.362
blockchain_technology_wikipedia.txt	wikipedia	0	0.5942	5.587	58.89	32.69	12.56	4.394
crispr_gene_editing_bard.txt	bard	1	0.5117	4.918	73.19	23.64	9.89	4.614
crispr_gene_editing_chatgpt.txt	chatgpt	1	0.6845	5.492	55.38	23.30	6.66	4.628
crispr_gene_editing_deepseek.txt	deepseek	1	0.5754	5.309	61.33	23.53	5.76	4.568
crispr_gene_editing_wikipedia.txt	wikipedia	0	0.6436	5.277	54.67	25.70	11.31	4.544
existentialism_bard.txt	bard	1	0.2834	6.384	30.99	26.80	7.52	4.213
existentialism_chatgpt.txt	chatgpt	1	0.3218	6.166	46.70	30.80	7.78	4.318
existentialism_deepseek.txt	deepseek	1	0.4223	5.899	50.87	29.38	6.36	4.268
existentialism_wikipedia.txt	wikipedia	0	0.6623	6.255	45.30	28.55	12.82	4.514
fall_of_roman_empire_bard.txt	bard	1	0.6178	5.853	44.47	21.13	6.80	4.417
fall_of_roman_empire_chatgpt.txt	chatgpt	1	0.5621	5.538	52.75	25.85	7.46	4.496
fall_of_roman_empire_deepseek.txt	deepseek	1	0.5522	5.128	59.47	31.78	11.00	4.402
fall_of_roman_empire_wikipedia.txt	wikipedia	0	0.5695	5.186	63.58	28.72	16.15	4.311
marvel_bard.txt	bard	1	0.5496	5.709	51.42	23.60	8.98	4.407
marvel_chatgpt.txt	chatgpt	1	0.5522	5.694	63.65	23.53	7.10	4.633
marvel_deepseek.txt	deepseek	1	0.5811	5.521	66.68	33.31	14.87	4.571
marvel_wikipedia.txt	wikipedia	0	0.5020	5.263	80.33	28.21	12.22	4.672
modernist_bard.txt	bard	1	0.4452	6.390	36.19	34.00	8.59	4.286
modernist_chatgpt.txt	chatgpt	1	0.5623	6.210	51.90	29.10	8.67	4.491
modernist_deepseek.txt	deepseek	1	0.6143	5.789	53.99	28.11	7.61	4.455
modernist_wikipedia.txt	wikipedia	0	0.3227	5.454	59.39	32.38	10.54	4.347
relativity_bard.txt	bard	1	0.5124	5.781	51.42	23.90	6.36	4.404
relativity_chatgpt.txt	chatgpt	1	0.5710	5.832	30.39	27.07	7.09	4.381
relativity_deepseek.txt	deepseek	1	0.5086	5.515	56.84	28.65	10.47	4.399
relativity_wikipedia.txt	wikipedia	0	0.5616	6.319	39.14	24.92	15.98	4.481