Beyond Perplexity: Character Distribution Signatures and the MDTA Benchmark for AI Text Detection

Priyadarshan Narayanasamy [*1] **Swastik Agrawal** [*1] **Klint Faber** [1] **Fardina Fathmiul Alam** [1]

Abstract

Training-free AI text detection methods primarily rely on model log-probabilities, achieving strong performance through approaches like Binoculars and DNA-DetectLLM. However, these methods face a fundamental ceiling as models are optimized through RLHF to produce human-like probability distributions. We introduce an alternative detection signal based on character distribution signatures. We provide theoretical foundations showing that AI models, trained on massive domain-balanced corpora, approximate global character patterns while humans exhibit domain-specialized distributions, creating a "Wall of Separation" where human-AI divergence significantly exceeds AI-AI divergence. To enable systematic evaluation, we construct the Models-Domains-Temperatures-Adversarials (MDTA) benchmark comprising 642,274 prompt-aligned samples across 4 models, 5 domains, 3 temperature settings, and 3 adversarial strategies, substantially expanding the HC3 dataset. We introduce the Letter Distribution Score (LD-Score), demonstrating low correlation ($r = 0.08$–0.13) with perplexity methods. When integrated with DNA-DetectLLM, Binoculars and FastDetectGPT via a non-linear classifier, LD-Score improves both AUROC and F1, with clearer gains in specialized domains where vocabulary constraints amplify the detection signal. The MDTA dataset can be accessed at: https://huggingface.co/datasets/nsp909/MDTA

1. Introduction

Large Language Models (LLMs) now generate content across countless domains, but distinguishing AI text from human writing has become critical in high-stakes applications. In academia, detection prevents plagiarism in student assignments and protects research integrity which is a pressing concern given that AI-generated manuscripts with fabricated citations have already infiltrated peer review at major conferences (Martin, 2025). Detection also guards against hallucinated information in legal documents and medical advice, where factual errors can have serious consequences.

Current state-of-the-art perplexity-based approaches, such as Binoculars (Hans et al., 2024) and DNA-DetectLLM (Zhu et al., 2025b), achieve strong detection performance by analyzing model log-probabilities. However, these methods face a fundamental ceiling as models are optimized to mimic human likelihood distributions, motivating complementary approaches that capture fundamental properties of text beyond perplexity.

We introduce an orthogonal detection signal based on letter distribution signatures. LLMs operate with comprehensive word probability distributions derived from extensive vocabularies and massive training corpora, while individual human writers exhibit specialized and skewed distributions with domain-specific patterns. This fundamental asymmetry causes LLMs to approximate global letter level statistical patterns, while humans deviate significantly through constrained vocabularies and stylistic preferences, creating a detectable signature at the letter level.

We introduce the Letter Distribution Score (LD-Score), an interpretable metric that quantifies letter distribution divergence. We provide mathematical foundations for why letter distributions differ systematically between human and AI text, then empirically validate the "Wall of Separation", revealing that letter distribution divergence between human and AI text noticeably exceeds divergence between different AI models.

Existing benchmark datasets present critical limitations: they either lack model-temperature diversity, domain coverage, prompt alignment, adversarial augmentation, or sufficient scale for robust evaluation. We contribute the Models-Domains-Temperatures-Adversarial (MDTA) benchmark, consisting of 642,274 prompt-aligned samples across 4 models, 5 domains, 3 temperature settings, and 3 adversarial strategies, substantially expanding the HC3 dataset (Guo

[*]Equal contribution [1]Department of Computer Science, University of Maryland, College Park, MD, USA. Correspondence to: Fardina Fathmiul Alam <fardina@umd.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

et al., 2023) with modern model responses, temperature variation, and targeted adversarial attacks including lipogrammatic constraints designed to challenge character-level detection methods.

We demonstrate that LD-Score provides an orthogonal detection signal to perplexity-based methods, exhibiting low correlation ($r = 0.08$–0.13) with existing approaches. Integration with state-of-the-art approaches (Binoculars, DNA-DetectLLM) yields consistent improvements in AUROC, and F1 scores, with the most notable gains observed in technical domains such as finance and medicine, where restricted vocabularies produce stronger character-level separation between human and AI text. Our approach requires no internal model access, providing a practical and effective complement to existing detection methods.

2. Related Work

AI-generated text detection has been an active research area since the early days of GPT-2, when Solaiman et al. (2019) developed the first major neural detector, a fine-tuned RoBERTa model. Since then, numerous deep learning approaches have emerged, with recent work reporting very high accuracies (95-99%) using fine-tuned BERT (Wang et al., 2024a), transformer architectures combined with linguistic features, and hybrid methods leveraging Bi-LSTM with attention mechanisms (Blake et al., 2025). Despite these impressive results in controlled settings, deep learning detectors face significant limitations. They struggle to generalize across different language models and domains, and are vulnerable to simple adversarial attacks (Sadasivan et al., 2025), with detectors trained on older models failing to detect outputs from newer models within the same generation or family. These limitations motivate the need for detection approaches that rely on more fundamental properties, both language-based and LLM-based, rather than learned model-specific patterns.

DetectGPT (Mitchell et al., 2023) established an important foundation for training-free detection by showing that text sampled from an LLM tends to occupy negative curvature regions of the model's log-probability function. Binoculars (Hans et al., 2024) introduced a contrastive framework that uses two models of different sizes to measure log-probability divergence, while DNA-DetectLLM (Zhu et al., 2025b) refined detection through probability perturbations across different model states and sampling strategies. More recently, BISCOPE (Guo et al., 2024) proposed a related but distinct logit-based framework that uses surrogate LLMs to extract bidirectional token-level features from forward and backward cross-entropy, rather than relying solely on standard next-token probability criteria. While these methods achieve strong empirical performance, they face fundamental limitations. As models are increasingly optimized through RLHF and constitutional AI to produce more human-like probability distributions, the probability gap narrows, creating a detection ceiling. Moreover, many of these approaches rely on access to model log-probabilities or logits, limiting their applicability in black-box settings. These limitations motivate the need for orthogonal detection signals that operate independently of model probability distributions and can augment existing methods.

Stylometric methods distinguish human from AI text by analyzing writing style features, such as Kumarage et al. (2023) who applied this approach to detect AI-generated tweets in social media timelines, Li & Zhang (2025) developed a comprehensive framework for Chinese social media with 34 features. These methods typically rely on features such as punctuation frequency, phraseology patterns, lexical complexity, sentence length statistics, linguistic diversity metrics, readability scores, and sentiment markers. However, these stylometric features represent surface-level characteristics rather than fundamental properties of text generation. They can be easily circumvented through fine-tuning, as models can be trained to mimic specific stylistic patterns once these features are identified by detectors.

N-gram frequency analysis represents a classical and popular approach to authorship attribution that has been applied to AI text detection. Gallé et al. (2021) demonstrated that repeated higher-order n-grams over-appear in machine-generated text, while Yang et al. (2023) proposed Divergent N-Gram Analysis (DNA-GPT), a training-free detection strategy that analyzes differences between original and regenerated text portions through n-gram analysis. Despite being well-established, n-gram methods face severe limitations for AI text detection: higher-order n-grams grow prohibitively sparse and computationally expensive, models can be explicitly trained to avoid or mimic specific n-gram patterns, and word-level analysis inherits the same lexical dependencies and adversarial vulnerabilities as stylometric approaches.

Drawing inspiration from naturally occurring statistical patterns like Benford's law (Benford, 1938) and Zipf's observations about language structure (Zipf, 1949), we hypothesize that letter-level distributions might encode detectable signatures based on the fundamental structure of language itself. Unlike word-level features that can be easily manipulated, letter distributions emerge from the aggregate effect of vocabulary selection and usage patterns, representing a more fundamental property of text generation. This intuition led us to develop the Letter Distribution Score as an orthogonal detection signal.

3. Letter Distribution Signatures

3.1. Theoretical Foundation: Exposure Scale and Domain Diversity

We formalize the fundamental asymmetry between human and AI text generation through their differential exposure to natural language and the resulting proximity to the global word probability distribution.

The Global Word Distribution. Let $P_{\text{global}}(w)$ denote the true word probability distribution across natural language—the population distribution aggregated over all contexts, domains, speakers, and time periods.

Convergence Through Exposure. When observing N words from a source, we estimate its word probability distribution $P(w)$. We treat the empirical word distribution as the marginal probability of observing each word. By the Law of Large Numbers, the empirical probability distribution converges to the true distribution with approximation error bounded by (Boucheron et al., 2013):

$$\|P(w) - P_{\text{global}}(w)\| = O\left(\frac{1}{\sqrt{N}}\right) \tag{1}$$

Larger exposure (N) yields better approximation, but only when samples are drawn proportionally from the target distribution.

AI Models: Massive, Domain-Balanced Exposure. Even the smallest of modern training corpora contain approximately 1.0 trillion tokens (Qiu et al., 2024; Shen et al., 2024). Using standard subword tokenization where one token ≈ 0.75 words, this yields $N_{\text{AI}} \approx 750$ billion words. These corpora span diverse domains: web text, Wikipedia, books, scientific papers, code, and conversational data. Under proportional sampling from P_{global}:

$$\|P_{\text{AI}}(w) - P_{\text{global}}(w)\| \approx \frac{1}{\sqrt{750 \times 10^9}} \approx 1.15 \times 10^{-6} \tag{2}$$

Humans: Limited, Domain-Specialized Exposure. Average adult reading speed is 238 words per minute (Brysbaert, 2019). Even assuming 8 hours daily (480 minutes) for 40 years:

$$N_{\text{human}} \approx 238 \times 480 \times 365 \times 40 \approx 1.67 \text{ billion words} \tag{3}$$

This represents an approximately 450-fold difference compared to AI training data. Under proportional sampling:

$$\|P_{\text{human}}(w) - P_{\text{global}}(w)\| \approx \frac{1}{\sqrt{1.67 \times 10^9}} \approx 2.4 \times 10^{-5} \tag{4}$$

approximately 21 times larger than AI model error.

The Domain Specialization Bias. The above assumes humans sample proportionally from P_{global} – an assumption that fails in practice. Humans exhibit strong domain specialization: medical professionals consume medical literature, engineers read technical documentation, and academics focus on scholarly work. Thus humans sample from domain-specific distributions $P_{\text{domain}} \neq P_{\text{global}}$.

Total divergence decomposes as:

$$\|P_{\text{human}}(w) - P_{\text{global}}(w)\| \approx \underbrace{\frac{1}{\sqrt{N_{\text{human}}}}}_{\text{statistical}} + \underbrace{\|P_{\text{domain}}(w) - P_{\text{global}}(w)\|}_{\text{domain bias}} \tag{5}$$

The domain bias is structural and persists regardless of reading volume. (We use "domain" flexibly to refer to any coherent grouping of text)

The Clustering Inequality. State-of-the-art models (GPT-5, Claude 4.5, Gemini 3.0) are increasingly trained on overlapping corpora—primarily Common Crawl, Wikipedia, books, and scientific papers. This shared training causes them to approximate P_{global} nearly identically:

$$P_{\text{GPT-5}}(w) \approx P_{\text{global}}(w) + \epsilon, \quad P_{\text{Claude}}(w) \approx P_{\text{global}}(w) + \epsilon' \tag{6}$$

where $\epsilon \approx \epsilon'$ due to correlated data. This yields:

$$\boxed{\max_{i,j \in \text{AI}} D(P_i, P_j) < \min_{\substack{h \in \text{human} \\ a \in \text{AI}}} D(P_h, P_a)} \tag{7}$$

where $D(\cdot, \cdot)$ denotes distributional divergence. This predicts that AI models form a tight cluster separated from human text: the "Wall of Separation".

3.2. From Word Distributions to letter Distributions

The above analysis establishes that $P_{\text{human}}(w)$ deviates from $P_{\text{global}}(w)$ due to domain specialization and limited exposure, while $P_{\text{AI}}(w) \approx P_{\text{global}}(w)$ due to massive and domain-balanced training. This divergence explains why word-level n-gram methods show empirical success. However, as discussed in Section 2, word-level features face critical limitations: extreme sparsity, signal dilution across high-dimensional space, and vulnerability to adversarial manipulation.

We address these limitations by projecting word distributions to letter-level statistics. Define the letter distribution as:

$$P(\ell) = \sum_{w \in \mathcal{V}} P(w) \frac{L(w, \ell)}{|w|} \tag{8}$$

where $\ell \in \{a, \ldots, z\}$, $L(w, \ell)$ counts occurrences of letter ℓ in word w, and $|w|$ is word length. This transformation

achieves approximately 1,900 times dimensionality reduction (50,000 to 26) while amplifying discriminative signals through aggregation.

The aggregation effect is key: multiple words with similar discriminative patterns contribute to the same letters, causing weak word-level signals to accumulate into stronger letter-level signals concentrated in just 26 dimensions.

Because the word-to-letter transformation is linear in $P(w)$, the clustering inequality established earlier in this section transfers directly from word to letter space:

$$\boxed{\max_{i,j \in \text{AI}} D(P_i^{\text{char}}, P_j^{\text{char}}) < \min_{\substack{h \in \text{human} \\ a \in \text{AI}}} D(P_h^{\text{char}}, P_a^{\text{char}})} \quad (9)$$

The "Wall of Separation" persists at the letter level, preserving the discriminative structure while eliminating the sparsity, computational cost, and adversarial vulnerability inherent at the word-level.

3.3. The Letter Distribution Score

We define the Letter Distribution Score (LD-Score) as a measure of letter-level distributional similarity between any two texts using a variation of Jensen-Shannon Distance (See Algorithm 1). The score quantifies how closely two texts match in their letter usage patterns. Lower scores indicate greater similarity and suggest the texts likely originate from the same source (both human or both AI), while higher scores indicate divergence.

3.4. Empirical Validation

We validate our theoretical predictions using the Ghostbuster dataset (Verma et al., 2024), analyzing LD-Score divergence across models and domains. For comprehensive domain-specific analysis and additional results, see Appendix B.

The Wall of Separation. Figure 1 presents pairwise LD-Scores in the essay domain. The matrix reveals the predicted two-scale structure: AI models cluster tightly with inter-model LD-Scores ranging from 0.0054 to 0.0217 (cool colors), while human text maintains consistently higher divergence from all AI models, with LD-Scores spanning 0.0250 to 0.0406 (warm colors).

Within the AI cluster, models with shared training backgrounds exhibit minimal LD-Score divergence: GPT-4 Omni and GPT-4 Turbo show LD-Score 0.0083, while Claude 3 Opus and Llama3 70B maintain 0.0054—the smallest AI-AI LD-Score observed. This validates our theoretical prediction that overlapping training corpora induce correlated approximations to $P_{\text{global}}(w)$.

Figure 2 provides a geometric visualization of the separation through PCA projection. The first two principal components capture 95.3% of total variance (PC1: 82.5%, PC2: 12.8%), indicating that letter distribution differences concentrate in a low-dimensional subspace. The spatial arrangement reinforces the idea that the distance from human to the nearest AI model substantially exceeds the maximum distance between any two AI models, geometrically confirming the clustering inequality established in the heatmap analysis.

Table 1. Correlation matrix between detection methods. LD-Score exhibits low correlation with perplexity-based methods (DNA-DetectLLM, Binoculars), indicating an orthogonal detection signal. In contrast, DNA and Binoculars are strongly correlated, reflecting shared likelihood-based modeling assumptions.

	DNA	Bino	Lexical	WD	LD
DNA	1.00	0.86	0.02	0.17	0.13
Bino	0.86	1.00	-0.02	0.10	0.08
Lexical	0.02	-0.02	1.00	0.22	0.11
WD-Score	0.17	0.10	0.22	1.00	0.42
LD-Score	0.13	0.08	0.11	0.42	1.00

Orthogonality to Existing Methods. To assess orthogonality with existing detection approaches, we compute Pearson correlations between detection signals. We compare: (1) DNA-DetectLLM score differences between AI and human text, (2) Binoculars score differences, (3) stylometric feature distance (punctuation frequency, sentence length, lexical diversity), (4) word-level RJSD (WD-Score), and (5) LD-Score. For the remainder of the paper, we refer to DNA-DetectLLM as **DNA** and Binoculars as **Bino**.

Table 1 presents the correlation matrix. As expected, DNA and Bino exhibit high correlation (r=0.86), reflecting their shared reliance on perplexity signals. LD-Score demonstrates low correlation with both perplexity-based methods (r=0.13 with DNA, r=0.08 with Bino), confirming near-orthogonality.

While stylometric features appear highly orthogonal (r $\approx$ -0.02-0.11), these superficial metrics are easily manipulated through post-processing and lack the robustness of distribution-based approaches, as discussed in Section 2. A comprehensive analysis of LD-score and 196 other stylometric methods is summarized in Appendix D.3, establishing LD-score as a valuable metric.

4. Dataset Construction

4.1. Motivation and Need

Existing benchmarks for AI-generated text detection each exhibit significant gaps. HC3 (Guo et al., 2023), while pioneering, covers only ChatGPT-3.5 and lacks multi-model and multi-temperature coverage. M4 (Wang et al., 2024b) spans multiple domains and models but uses inconsistent model sets across domains and omits temperature variation. Ghostbuster (Verma et al., 2024) offers prompt-aligned

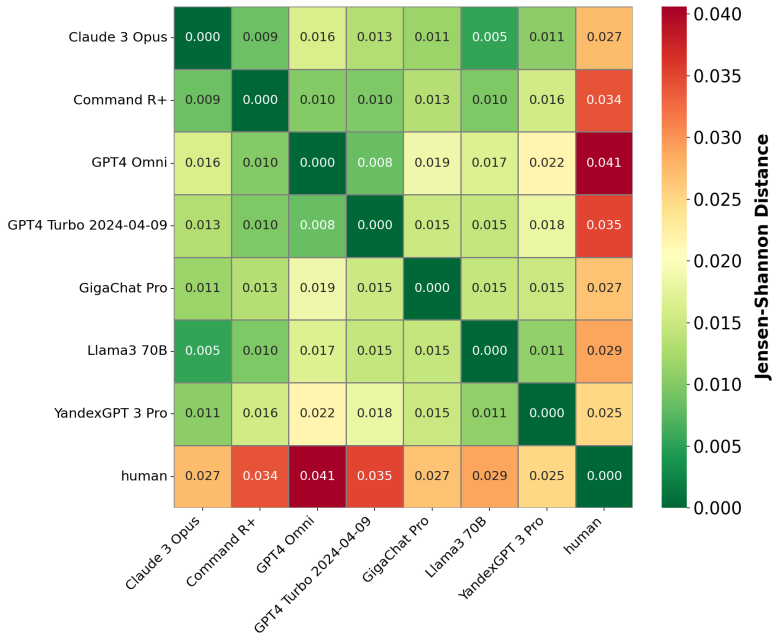


Figure 1. Pairwise LD-Scores between complete (A-Z) letter distributions in the Essay domain. The human row (bottom) exhibits systematically higher divergence from all AI models compared to AI-to-AI distances.

multi-model responses but is limited to ∼1,000 samples per domain at a single temperature. RealDet (Zhu et al., 2025a) improves scale and breadth but similarly lacks consistent cross-domain model coverage, temperature variation, and relies on relatively weak adversarial attacks. We discuss existing datasets in-depth in Appendix C.

These datasets suffer from one or more of the following limitations: (1) reliance on outdated models, (2) lack of prompt-level alignment across generators, (3) absence of temperature variation, (4) limited domain coverage, (5) insufficient scale for robust statistical analysis, or (6) lack of targeted adversarial augmentation beyond standard paraphrasing. We address all of these concerns with our comprehensive dataset construction, which comprises **642,274** samples spanning five domains, four models, three temperature settings, and three adversarial attack strategies.

4.2. Dataset Generation

Base Dataset and Human Text. We used the HC3 corpus (Guo et al., 2023) as our foundation, leveraging its 24,322 unique prompts spanning five domains with 58,546 authentic human responses. This provides crucial domain diversity, ranging from highly technical fields (finance, medicine) to conversational contexts (Reddit_ELI5), with average response lengths varying from 186.8 to 1,301.6 tokens.

AI Model Response Generation. We generated synthetic responses using four recent mid-sized state-of-the-art open-source models: **Llama 3.1 8B**, **Gemma 3 12B**, **Qwen2.5-VL 7B**, and **Ministral 8B**. These models represent diverse architectural approaches and training paradigms while remaining computationally accessible for reproduction.

For each of the 24,322 prompts, we generated three responses per model at temperature settings of **0.2** (deterministic), **0.5** (balanced), and **0.8** (stochastic). This temperature stratification is essential, as sampling temperature directly modulates vocabulary probability distributions. An overview of the dataset composition is tabulated in Table 4 in Appendix C.

4.3. Adversarial Augmentation

To evaluate robustness against evasion, we augmented the dataset with adversarial variants generated by the originating model itself: (A) a standard paraphrase, (B) a paraphrase avoiding a randomly selected letter ℓ_1, and (C) a paraphrase avoiding two distinct letters $\ell_1 \neq \ell_2$. These constraints directly stress-test letter-distribution-based detection by forcing shifts in character-level statistics, and effectively double the dataset size. Attack success analysis is presented in Appendix C.

5. Implementation for LLM Text Detection

Having discussed the empirical results proving the "Wall of Separation" between letter distributions of Human and

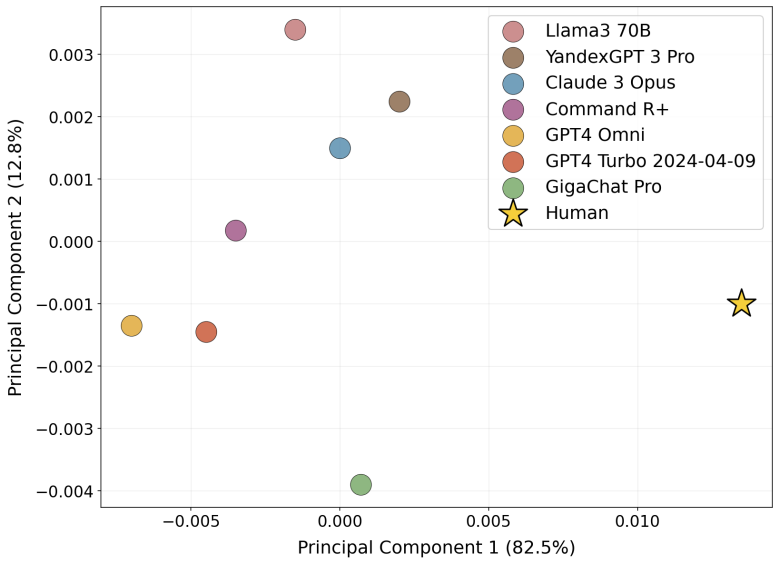


Figure 2. Principal Component Analysis (PCA) of letter probability distributions shows the human letter probability distribution is clearly distinct from all AI models, which cluster separately, indicating a consistent distributional shift in AI-generated text.

LLM-generated text, we now develop an approach utilizing this finding and set up experiments to analyze improvement in performance of existing training-free black-box methods (**DNA**, **Bino** and **FastDetectGPT** (Bao et al., 2024)) through an augmentation strategy.

We consider responses from 3 candidate models to adequately quantify the difference in letter distributions between input text and that of reference AI text. Our scoring mechanism involves combining the text outputs from all 3 candidate models and computing the LD-score (Subsection 3.3) between the input text and this combined reference distribution.

To retain the full discriminative power of both signals, we instead represent each sample s as a two-dimensional feature vector:

$$\mathbf{x}(s) = \begin{bmatrix} R(s) \\ \text{LD-Score}(P_{\text{test}}(s) \parallel P_{\mathcal{M}}(s)) \end{bmatrix} \tag{10}$$

where $R(s)$ denotes the base detector score for sample s, P_{test} is the letter distribution of the input text, and $P_{\mathcal{M}}$ is the letter distribution computed over the combined text outputs of all candidate models in $\mathcal{M}$.

We employ a Support Vector Machine (SVM) with a radial basis function (RBF) kernel over our 2-dimensional input vector $x(s)$(discussed further in Appendix A.2).

6. Experiments

6.1. Experimental Setup

Datasets. We set up our experiments by sourcing our data from the MDTA and Ghostbuster datasets, drawing AI samples from 4 different models. For Ghostbuster, we use the following models: "GPT4 Turbo 2024-04-09", "GPT4 Omni", "Claude 3 Opus", and "GigaChat Pro". We then separate each dataset into 4 different class-balanced sections, one for each of the 4 models serving as the "AI model". The remaining model responses form the three models we draw LD-score from.

Baselines DNA, Bino and FastDetectGPT (FDGPT) are adopted as baseline training-free methods and are augmented with the LD-Score. "Falcon-7b-Instruct" and "Falcon-7b" serve as the reference (performer) and observer models, respectively, in DNA and Bino approaches. FDGPT uses GPT-J-6B for sampling and GPT-Neo-2.7B for scoring.

6.2. Performance Analysis

Table 2 reports F1 and AUROC for baselines and their LD-Score augmentations with only 100 training samples (50 human, 50 AI). LD-DNA achieves the highest average F1 (0.94) and AUROC (0.97), with consistent improvements over DNA across nearly all domains. The gains are largest in structured domains, Reuters (ΔF1: +0.02), Reddit ELI5 (ΔF1: +0.04), and Open QA (ΔF1: +0.03), where domain-specialized vocabularies amplify the LD signal, consistent with the theoretical framework in Section 3. Finance and Wiki CSAI show minimal gains, as their broader vocabu-

laries weaken distributional separation. Notably, LD-Score augmentation consistently outperforms perplexity-only ensembles (DNA+Bino, DNA+FastDetectGPT), demonstrating that the orthogonality of LD-Score to perplexity signals provides complementary discriminative power beyond what combining perplexity methods alone achieves. Across all methods, augmentation with LD-Score also reduces variance, indicating more stable detection under limited training data.

Unbalanced Training Regime To further stress-test LD-Score fusion, we evaluate under a more challenging unbalanced training regime where human samples are scarce (Table 3). In this setting, DNA and Binoculars struggle to learn a reliable decision boundary, improving only gradually with more data (DNA: $0.847 \rightarrow 0.886$). By contrast, their LD-Score-augmented counterparts converge rapidly from the outset, achieving AUROC of 0.935 and 0.918 with only 100 samples, gains of $+0.088$ and $+0.122$ respectively. Augmented methods plateau early and maintain their advantage throughout, with the diminishing Δ at larger sample sizes reflecting baselines slowly catching up via improved threshold calibration rather than any degradation of the fusion. This is particularly significant for real-world AI detection, where human-generated text is harder to collect and label at scale. Furthermore, these results reveal that the LD-Score requires surprisingly few domain examples to capture a domain's letter distribution, demonstrating that domain specialization does not pose a practical obstacle to deployment.

Adversarial Experiments. The adversarial variants tested here represent a stress test of our augmentation pipeline, since the LD-Score is derived from clean, non-adversarial model responses and is never optimized against these attacks; full results are shown in the Appendix E. Despite this, LD-Score augmentation improves AUROC over the base detectors in nearly every condition, matching or exceeding base performance in 29 of 30 domain-attack comparisons. Gains are consistent across all three attack types, with average AUROC improvements of $+0.005$ and $+0.010$ for DNA and Binoculars respectively under paraphrase attacks (A), $+0.006$ and $+0.014$ under single letter removal (B), and $+0.006$ and $+0.012$ under dual letter removal (C). Although F1 score improvements are less consistent than AUROC gains, the augmented variants remain better than or broadly comparable to their unaugmented counterparts.

7. Conclusion and Limitations

This work introduces letter distribution signatures as an orthogonal detection signal for AI-generated text, establishing a "Wall of Separation" where human-AI divergence systematically exceeds AI-AI divergence. Integration with state-of-the-art methods yields consistent AUROC and F1 improvements, with LD-DNA achieving average F1 of 0.921 and AUROC of 0.960 compared to 0.907 and 0.946 for DNA alone, and low correlation ($r = 0.08 - 0.13$) with perplexity-based methods confirming orthogonality. Key limitations include domain dependence, where the signal is strongest in specialized domains and weaker in open-domain settings, and reliance on surrogate LLMs, which introduces computational overhead. Future work should explore stronger fusion strategies, more comprehensive adversarial robustness including persona prompting and sophisticated paraphrasing attacks, training data contamination detection via distribution matching, and extensions to AI image detection through spectral distribution analysis.

Acknowledgements

The authors acknowledge the resources provided by High-Performance Computing at the University of Maryland (https://hpcc.umd.edu/) for conducting the research reported in this paper.

Impact Statement

This work studies statistical regularities in character-level distributions to distinguish human-written text from AI-generated text in a training-free and model-agnostic manner. The proposed approach may support applications in academic integrity, misinformation analysis, and content provenance by providing lightweight signals that complement existing detection methods without requiring access to model internals or proprietary likelihoods. By focusing on aggregated lexical structure rather than semantic content, the method is computationally efficient and broadly applicable across domains and models. At the same time, AI text detection carries important limitations and ethical considerations. No detection method is perfectly reliable, and false positives may disproportionately affect non-native speakers, domain specialists, or stylistically unconventional human writers. Our results should therefore be interpreted as providing probabilistic signals rather than definitive judgments. The proposed approach is not intended to be used in isolation for high-stakes decisions, but rather as one component in a broader assessment framework.

We view this work primarily as a contribution toward understanding structural properties of AI-generated language rather than as a definitive solution to AI text attribution. Future work should further study robustness, fairness across populations, and appropriate human oversight when deploying such methods in real-world settings.

Table 2. Detection performance with 100 balanced training samples (50 AI + 50 human) for SVM training and threshold calibration. The Ghostbuster column reports the average AUROC/F1 across Essay, Reuters, and WP domains. LD-X denotes augmentation of method X with the LD-Score via RBF-SVM fusion. Results are mean ± std over 5 runs with different seeds. **Bold** indicates best per column.

	Ghostbuster	MDTA					
Method	Avg	Finance	Medicine	Open QA	Reddit ELI5	Wiki CSAI	Avg
AUROC							
DNA	0.994±0.005	0.958±0.016	0.996±0.005	0.812±0.043	0.971±0.008	**0.992**±0.006	0.946±0.021
Bino	0.985±0.008	0.946±0.014	0.995±0.006	0.817±0.042	0.966±0.009	0.990±0.008	0.943±0.021
FDGPT	0.935±0.019	0.871±0.011	0.954±0.011	0.774±0.048	0.930±0.011	0.963±0.011	0.898±0.024
LD-DNA	**0.995**±0.005	**0.974**±0.008	**0.998**±0.001	0.856±0.029	**0.982**±0.007	**0.992**±0.004	**0.960**±0.014
LD-Bino	0.989±0.005	0.970±0.006	0.997±0.003	**0.859**±0.031	0.978±0.008	0.990±0.005	0.959±0.015
LD-FDGPT	0.960±0.016	0.888±0.014	0.961±0.010	0.808±0.026	0.934±0.015	0.956±0.011	0.909±0.016
DNA+Bino	0.993±0.005	**0.974**±0.008	0.994±0.004	0.832±0.039	0.958±0.009	0.989±0.011	0.949±0.019
DNA+FDGPT	**0.995**±0.004	0.971±0.010	0.996±0.003	0.826±0.038	0.965±0.010	0.990±0.006	0.950±0.018
F1 Score							
DNA	0.971±0.014	**0.933**±0.011	0.976±0.009	0.756±0.040	0.913±0.026	**0.958**±0.014	0.907±0.023
Bino	0.943±0.027	0.905±0.019	0.977±0.009	0.752±0.048	0.895±0.024	0.949±0.017	0.896±0.027
FDGPT	0.864±0.032	0.751±0.043	0.879±0.028	0.701±0.042	0.832±0.035	0.896±0.025	0.812±0.035
LD-DNA	**0.980**±0.007	0.929±0.017	**0.982**±0.006	**0.788**±0.027	**0.949**±0.012	0.955±0.010	**0.921**±0.016
LD-Bino	0.965±0.013	0.918±0.017	0.976±0.008	0.784±0.028	0.936±0.013	0.956±0.008	0.914±0.017
LD-FDGPT	0.916±0.022	0.770±0.032	0.907±0.013	0.720±0.027	0.856±0.015	0.885±0.018	0.828±0.022
DNA+Bino	0.971±0.008	0.927±0.015	0.979±0.005	0.764±0.044	0.911±0.018	**0.958**±0.011	0.908±0.023
DNA+FDGPT	0.976±0.010	0.925±0.017	0.978±0.006	0.757±0.052	0.914±0.019	0.957±0.011	0.906±0.026

Table 3. AUROC under an unbalanced training regime (human samples scarce) as a function of training set size. DNA and Binoculars improve only gradually with more data, while their LD-Score-augmented counterparts (LD-DNA, LD-Bino) converge rapidly and maintain a consistent advantage throughout.

Train	DNA	LD-DNA	Δ	Bino	LD-Bino	Δ
100	0.847	0.935	**+0.088**	0.796	0.918	**+0.122**
250	0.863	0.949	**+0.086**	0.823	0.945	**+0.122**
500	0.864	0.945	**+0.081**	0.812	0.942	**+0.130**
1000	0.886	0.945	**+0.059**	0.801	0.935	**+0.134**

References

Bao, G., Zhao, Y., Teng, Z., Yang, L., and Zhang, Y. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, 2024. URL https://arxiv.org/abs/2310.05130.

Benford, F. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

Blake, J., Miah, A. S. M., Kredens, K., and Shin, J. Detection of ai-generated texts: A bi-lstm and attention-based approach. *IEEE Access*, 13:71563–71576, 2025. doi: 10.1109/ACCESS.2025.3562750.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Brysbaert, M. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047, 2019. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2019.104047. URL https://www.sciencedirect.com/science/article/pii/S0749596X19300786.

Endres, D. M. and Schindelin, J. E. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Flesch, R. A new readability yardstick. *Journal of Applied*

Psychology, 32(3):221–233, 1948. doi: 10.1037/h0057532.

Gallé, M., Rozen, J., Kruszewski, G., and Elsahar, H. Unsupervised and distributional detection of machine-generated text, 2021. URL https://arxiv.org/abs/2111.02878.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*, 2023.

Guo, H., Cheng, S., Jin, X., Zhang, Z., Zhang, K., Tao, G., Shen, G., and Zhang, X. Biscope: Ai-generated text detection by checking memorization of preceding tokens. *Advances in Neural Information Processing Systems 37*, 2024. URL https://api.semanticscholar.org/CorpusID:276288166.

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., and Goldstein, T. Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. URL https://arxiv.org/abs/2401.12070.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951. ISSN 00034851. URL http://www.jstor.org/stable/2236703.

Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., and Liu, H. Stylometric detection of ai-generated text in twitter timelines, 2023. URL https://arxiv.org/abs/2303.03697.

Li, Z. and Zhang, Q. Linguistic differences between ai and human comments in weibo: Detect ai-generated text through stylometric features. pp. 31–42. 2025. ISBN 978-981-95-2724-3. doi: 10.1007/978-981-95-2725-0_3.

Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Martin, S. Startup investigation reveals 50 peer-reviewed papers contained ai-hallucinated citations, December 2025. URL https://betakit.com/start-up-investigation-reveals-50-peer-reviewed-papers-contained-hallucinated-citations/.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023. URL https://arxiv.org/abs/2301.11305.

Przystalski, K., Argasiński, J. K., Grabska-Gradzińska, I., and Ochab, J. K. Stylometry recognizes human and llm-generated texts in short samples. *Expert Systems with Applications*, 296:129001, January 2026. ISSN 0957-4174. doi: 10.1016/j.eswa.2025.129001. URL http://dx.doi.org/10.1016/j.eswa.2025.129001.

Qiu, J., Lv, H., Jin, Z., Wang, R., Ning, W., Yu, J., Zhang, C., Li, Z., Chu, P., Qu, Y., Shi, J., Lu, L., Peng, R., Zeng, Z., Tang, H., Lei, Z., Hong, J., Chen, K., Fei, Z., Xu, R., Li, W., Tu, Z., Dahua, L., Qiao, Y., Yan, H., and He, C. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset, 2024. URL https://arxiv.org/abs/2402.19282.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can ai-generated text be reliably detected?, 2025. URL https://arxiv.org/abs/2303.11156.

Shen, Z., Tao, T., Ma, L., Neiswanger, W., Liu, Z., Wang, H., Tan, B., Hestness, J., Vassilieva, N., Soboleva, D., and Xing, E. Slimpajama-dc: Understanding data combinations for llm training, 2024. URL https://arxiv.org/abs/2309.10818.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. Release strategies and the social impacts of language models, 2019. URL https://arxiv.org/abs/1908.09203.

Verma, V., Fleisig, E., Tomlin, N., and Klein, D. Ghostbuster: Detecting text ghostwritten by large language models, 2024. URL https://arxiv.org/abs/2305.15047.

Wang, H., Li, J., and Li, Z. Ai-generated text detection and classification based on bert deep learning algorithm, 2024a. URL https://arxiv.org/abs/2405.16422.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O. M., Mahmoud, T., Sasaki, T., Arnold, T., Aji, A. F., Habash, N., Gurevych, I., and Nakov, P. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, 2024b. URL https://arxiv.org/abs/2305.14902.

Wu, J., Zhan, R., Wong, D. F., Yang, S., Yang, X., Yuan, Y., and Chao, L. S. Detectrl: Benchmarking llm-generated text detection in real-world scenarios, 2025. URL https://arxiv.org/abs/2410.23746.

Yang, X., Cheng, W., Wu, Y., Petzold, L., Wang, W. Y., and Chen, H. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text, 2023. URL https://arxiv.org/abs/2305.17359.

Zhu, X., Ren, Y., Cao, Y., Lin, X., Fang, F., and Li, Y. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction, 2025a. URL https://arxiv.org/abs/2505.05084.

Zhu, X., Ren, Y., Fang, F., Tan, Q., Wang, S., and Cao, Y. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm, 2025b. URL https://arxiv.org/abs/2509.15550.

Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA, 1949.

Algorithm 1 Letter Distribution Score Computation

input Two text samples T_1 and T_2
output LD-Score between T_1 and T_2

// Extract letter distributions
for each text $t \in \{T_1, T_2\}$ **do**
  for each letter $\ell \in \{a, \ldots, z\}$ **do**
    $P_t(\ell) \leftarrow \frac{\text{count}(\ell \text{ in } t)}{\sum_{\ell'} \text{count}(\ell' \text{ in } t)}$
  end for
end for

// Compute Jensen-Shannon Divergence
$M \leftarrow \frac{1}{2}(P_{T_1} + P_{T_2})$
$\text{KL}(P\|Q) \leftarrow \sum_{i=1}^{26} P(i) \log \frac{P(i)}{Q(i)}$ {Kullback-Leibler divergence (Kullback & Leibler, 1951)}
$\text{JSD}(P_{T_1}\|P_{T_2}) \leftarrow \frac{1}{2}[\text{KL}(P_{T_1}\|M) + \text{KL}(P_{T_2}\|M)]$ {Jensen-Shannon divergence (Lin, 1991)}

// Compute Root Jensen-Shannon Distance **return** LD-Score$(T_1, T_2) = \sqrt{\text{JSD}(P_{T_1}\|P_{T_2})}$ {RJSD metric (Endres & Schindelin, 2003)}

A. Methodology Details

A.1. Algorithm

Algorithm 1 details the computation of the Letter Distribution Score between two text samples. Given texts T_1 and T_2, we first extract normalized letter frequency probability distributions P_{T_1} and P_{T_2} over the 26-character English alphabet. We then compute the Jensen-Shannon Divergence (JSD) between these distributions, a symmetric and smoothed variant of the Kullback-Leibler divergence that avoids the numerical instability arising when one distribution assigns zero probability to a letter. Taking the square root yields the Root Jensen-Shannon Distance (RJSD), which satisfies the triangle inequality and thus constitutes a proper metric space—a desirable formal property that raw JSD does not provide. The resulting LD-Score is bounded in $[0, 1]$, where values near zero indicate near-identical letter distributions and values approaching one indicate maximally divergent distributions.

A.2. SVM Implementation

The SVM operates using a radial basis function (RBF) kernel to capture the non-linear decision boundary:

$$\hat{y}(s) = \text{sign}\left(\sum_i \alpha_i y_i \, K(\mathbf{x}_i, \mathbf{x}(s)) + b\right) \tag{11}$$

where $K(\mathbf{x}_i, \mathbf{x}(s)) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}(s)\|^2\right)$ is the RBF kernel.

This formulation allows the classifier to learn curved, non-linear boundaries in the two-dimensional feature space, effectively exploiting the complementary structure of both detection signals. While we use DNA and Binoculars as baselines here, this approach generalizes to any perplexity-based method paired with the LD-score.

B. Ghostbuster Analysis

B.1. Domain-Specific Dataset Analysis

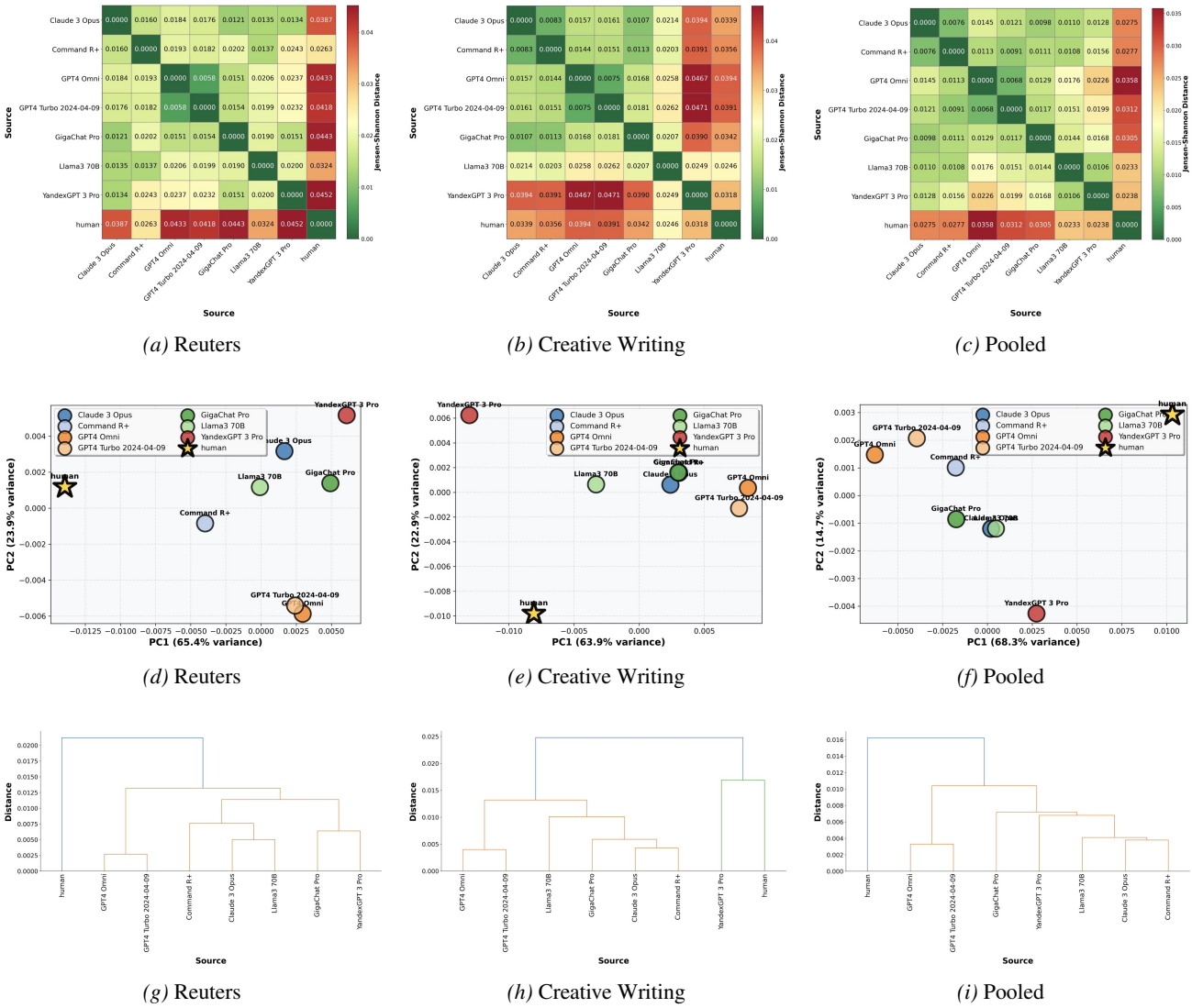
This appendix provides detailed pairwise Jensen-Shannon distance matrices and hierarchical clustering dendrograms for individual domains, complementing the essay domain analysis presented in the main text (Section 3.4).

The hierarchical clustering dendrograms (bottom row of Figure 3) reveal the predicted clustering structure. Across all domains, similar models cluster together, with GPT-4 family models showing particularly tight grouping. Critically, human text mostly splits at the top level of the hierarchy, forming a completely separate cluster from all AI models. This top-level split validates our theoretical prediction that human text occupies a fundamentally distinct region in letter distribution space, separated by the Wall of Separation from the AI model cluster.

B.2. Domain-Dependent Tightness of the AI Cluster

Figure 4 illustrates letter-level log-probability deviations from the human baseline across two domains: Essays and Writing Prompts. In structured domains such as essays, AI-generated texts exhibit highly consistent letter-level deviations across models, resulting in tightly overlapping curves. This reflects strong AI–AI clustering: constrained task structure, formal tone, and standardized vocabulary encourage all models to sample similarly from their shared approximation of the global language distribution.

In contrast, creative domains such as writing prompts induce substantially higher variability across AI models. Open-ended generation amplifies stylistic choices, narrative voice, and lexical experimentation, increasing divergence both among AI models and relative to the human baseline. This expansion of the AI cluster reduces the AI–AI similarity margin and weakens separation, consistent with the domain bias term in Eq. 5.



(a) Reuters *(b)* Creative Writing *(c)* Pooled

(d) Reuters *(e)* Creative Writing *(f)* Pooled

(g) Reuters *(h)* Creative Writing *(i)* Pooled

Figure 3. Domain-specific LD-Score analysis on Ghostbuster dataset. Top row: Pairwise Jensen-Shannon distance matrices. Middle row: PCA projections visualizing geometric separation. Bottom row: Hierarchical clustering dendrograms. Reuters (specialized news) shows strongest separation, Creative Writing (general, unstructured) shows weakest separation, and Pooled results demonstrate overall robustness.

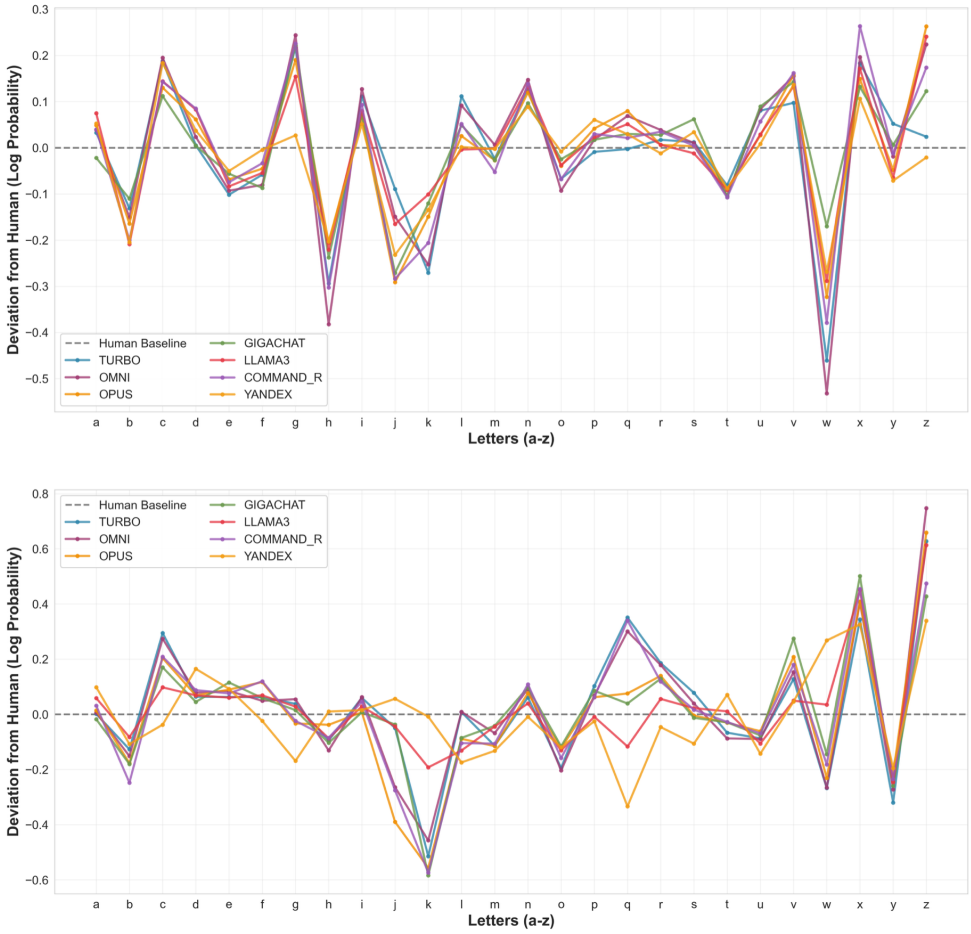


Figure 4. Letter-level log-probability deviations from the human baseline across two domains: Essays (top) and Writing Prompts (bottom). Curves correspond to different AI models. In structured domains such as essays, AI models exhibit tightly clustered letter distributions, indicating strong AI–AI similarity. In contrast, open-ended creative writing prompts induce greater variability across models, illustrating how domain restrictiveness controls the tightness of the AI cluster and the strength of separation from human text.

Table 4. Dataset composition by domain in the MDTA benchmark. All AI models possess the same number of samples at each temperature. Adversarial variants (Paraphrase, Avoid ℓ_1, Avoid ℓ_1 & ℓ_2) are generated from the $t = 0.5$ responses using the originating model. Reddit_ELI5 has a particularly large number of human samples because the MDTA dataset contains 3 human responses per prompt.

		Standard			Adversarial		
Domain	Human	t=0.2	t=0.5	t=0.8	Para.	Avoid ℓ_1	Avoid ℓ_1 & ℓ_2
Finance	3,933	15,732	15,732	15,732	15,732	15,732	15,732
Medicine	1,248	4,992	4,992	4,992	4,992	4,992	4,992
Open QA	1,187	4,748	4,748	4,748	4,748	4,748	4,748
Reddit ELI5	51,336	68,448	68,448	68,448	68,448	68,448	68,448
Wiki CSAI	842	3,368	3,368	3,368	3,368	3,368	3,368
Total	**58,546**	**97,288**	**97,288**	**97,288**	**97,288**	**97,288**	**97,288**

C. The MDTA Dataset

C.1. Existing Datasets

Current datasets for AI-generated text detection suffer from critical limitations that hinder the development of robust, generalizable detection methods. The original HC3 (Human ChatGPT Comparison Corpus) dataset (Guo et al., 2023), while pioneering in providing domain-diverse human-AI text pairs, contains only ChatGPT-3.5 responses paired with human text. Given the rapid advancement in language model capabilities since early 2023, this dataset is now outdated and fails to capture the linguistic characteristics of modern state-of-the-art models. More fundamentally, it lacks the multi-model and multi-temperature coverage necessary for developing detection methods that generalize across different AI systems and generation strategies.

Beyond HC3, existing benchmark datasets exhibit complementary but insufficient coverage for comprehensive distributional analysis. The M4 benchmark dataset (Wang et al., 2024b) is one of the most comprehensive efforts to date, spanning multiple domains and including responses from a wide range of language models. However, it does not consistently employ the same set of models across all domains, which prevents proper cross-domain comparisons of model-specific characteristics. More critically, the models included in M4 are largely outdated, having since been succeeded by substantially more capable language models on both the proprietary and open-source fronts. Additionally, M4 lacks temperature variation in its generation strategy and its adversarial attack configurations, while valuable at the time of release, have become less representative of modern evasion techniques.

The Ghostbuster dataset (Verma et al., 2024) addresses several of these shortcomings, providing prompt-aligned multi-model responses across three domains: creative writing (Writing Prompts), news (Reuters), and student essays. However, it provides only approximately 1,000 samples per domain and generates all responses at a single temperature setting. This limits the dataset's utility for studying how decoding stochasticity influences the statistical properties of generated text more broadly, as low temperatures produce deterministic, repetitive outputs while high temperatures yield greater lexical and structural variation.

More recently, RealDet (Zhu et al., 2025a) introduced a large and comprehensive benchmark spanning many domains, prompts, and LLMs, representing an important step forward in dataset scale and breadth. However, it is less suitable for controlled distributional comparison. The dataset does not consistently provide the same set of LLMs across all domains, which makes systematic cross-domain, cross-model comparisons more difficult. In addition, many of the included models have since been surpassed by substantially stronger proprietary and open-source systems, limiting its usefulness as a benchmark for studying the behavior of current-generation LLMs. RealDet also does not explicitly vary generation temperature, preventing analysis of how decoding stochasticity affects textual distributions. Finally, although it includes adversarial attacks, these largely rely on relatively standard paraphrasing and token-level perturbations, which are weaker than more modern adaptive attacks in which the source LLM itself is instructed to rewrite text under targeted lexical constraints.

C.2. Domain-Specific Analysis

Table 4 summarizes the composition of the entire MDTA dataset, with Table 5 detailing the average word and character counts per sample across domains. Figure 5 presents domain-specific LD-Score analysis across these domains, extending our earlier Ghostbuster findings to a larger and more diverse benchmark with additional models and temperature variations.

Table 5. MDTA benchmark dataset statistics across domains.

Domain	Avg Words	Avg Chars
Finance	145	947
Medicine	191	1,120
Open QA	45	268
Reddit ELI5	160	891
Wiki CSAI	123	916

Figure 5 presents analogous analysis on the MDTA benchmark dataset, which includes additional models and temperature variations.

The MDTA results (Figure 5) corroborate findings from the Ghostbuster dataset. Hierarchical clustering consistently shows human text splitting at the top level across all domains. Similar models cluster together within the AI group, with the dendrogram structure reflecting training data overlap. Notably, specialized domains (Finance, Medicine, Reddit ELI5) display stronger separation and clearer hierarchical structure compared to general domains (Wiki CSAI), validating that domain specialization amplifies the detection signal as predicted by our theoretical framework.

C.3. Adversarial Dataset Analysis (MDTA)

As mentioned earlier, for each AI-generated response at temperature 0.5, we used the originating model itself to produce three adversarial rewrites: (A) a standard paraphrase, (B) a paraphrase avoiding a randomly selected letter ℓ_1, and (C) a paraphrase simultaneously avoiding two letters $\ell_1 \neq \ell_2$. We perform analysis. In Figures 6 7 8, we analyze the effectiveness of letter-removal attacks (B) and (C) in altering the responses of the models.

Overall, 100% avoidance rates are low across all models (Figure 7), with the best-performing model (`ministral-8b`) achieving full avoidance in only $\sim$10% of samples. As shown in Figure 8, success is heavily skewed toward rare letters—`Z`, `J`, `Q`, and `X` are avoided in 30–45% of samples, while high-frequency letters like `T`, `S`, and `N` are nearly impossible to avoid. This means adversarial character-distribution manipulation succeeds precisely where it matters least for detection: rare letters contribute little to the LD-Score's discriminative signal, leaving the detector's core features largely intact.

C.4. Linguistic Complexity Analysis

We assess text complexity through readability, lexical diversity, and n-gram analysis.

n-gram Analysis. We further perform n-gram analysis by examining frequency distributions of contiguous word sequences, where an n-gram consists (uni and bi) of n consecutive words. According to Table 6, it is clearly evident that the cumulative human unigram-vocabulary is higher than any of the AI models. This is in accordance to Subsection 3.1 (Convergence through Exposure) since the human responses are sourced from various different speakers.

Readability. Readability is quantified using the Flesch–Kincaid Grade Level (FKGL) (Flesch, 1948), defined as

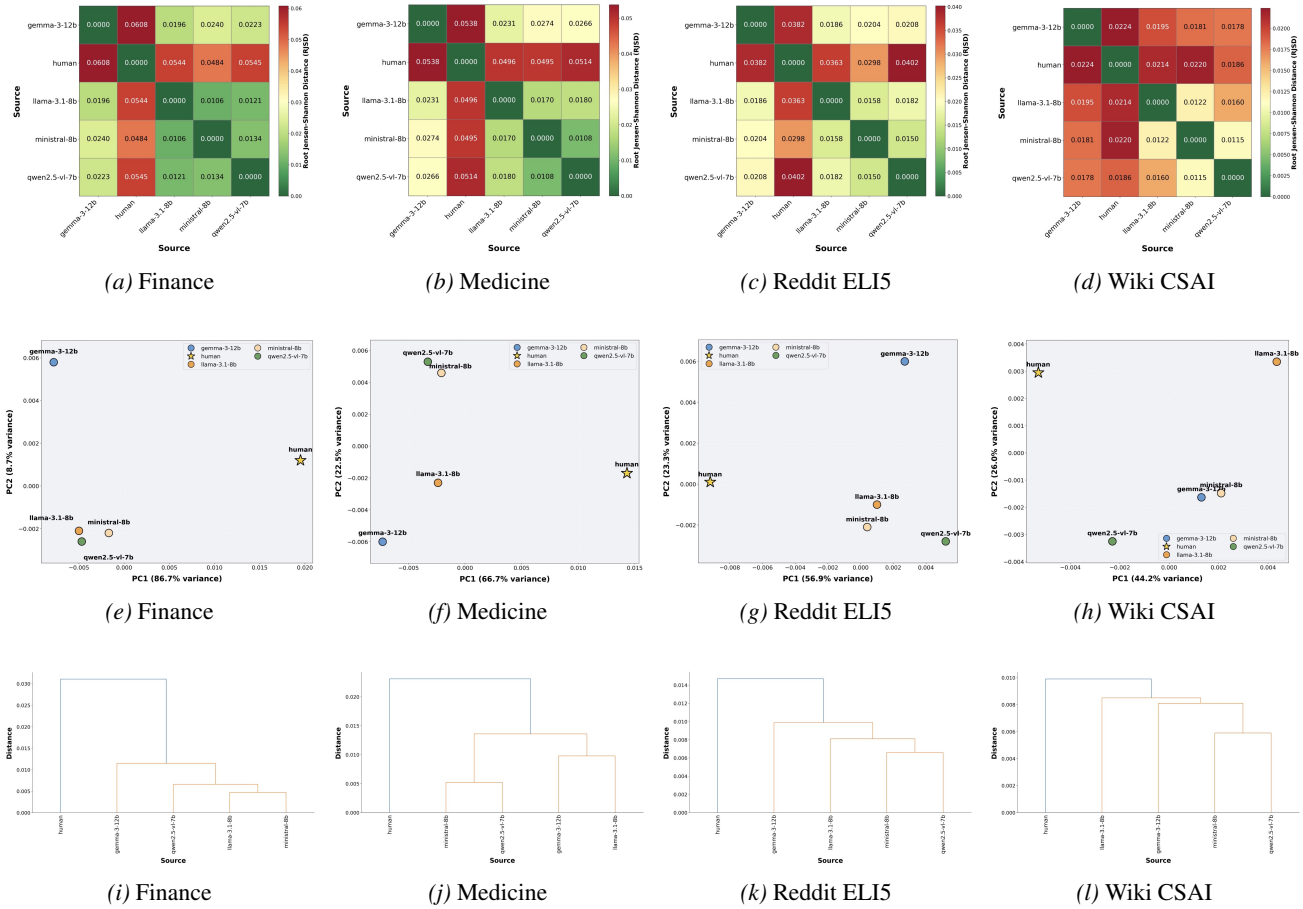
$$\text{FKGL} = 0.39 \cdot \frac{W}{S} + 11.8 \cdot \frac{Sy}{W} - 15.59, \tag{12}$$

where W denotes the number of words, S the number of sentences, and Sy the number of syllables. FKGL estimates the U.S. grade level required to comprehend the text. The results, as depicted in Fig. 9, are expected. A specific human tends to explain simply, using smaller and relatively simpler words than language models. This is amplified if we only visualize the Reddit_ELI5 domain. While n-gram analysis is performed considering the entire human corpus within the dataset, readability analysis is drawn per sample.

Lexical Diversity. We calculate the lexical diversity score (LDS) as defined in (Wu et al., 2025).

$$\text{LDS} = \frac{|V|}{N}, \tag{13}$$

where $|V|$ is the number of unique word types and N is the total number of word in the input text. Higher LDS values indicate richer vocabulary usage, while lower values suggest more repetitive language. We observe similar results: a



(a) Finance *(b)* Medicine *(c)* Reddit ELI5 *(d)* Wiki CSAI

(e) Finance *(f)* Medicine *(g)* Reddit ELI5 *(h)* Wiki CSAI

(i) Finance *(j)* Medicine *(k)* Reddit ELI5 *(l)* Wiki CSAI

Figure 5. Domain-specific LD-Score analysis on some domains of the MDTA dataset. Top row: Pairwise distance matrices. Middle row: PCA projections. Bottom row: Hierarchical clustering dendrograms. Finance and Medicine (specialized domains) exhibit stronger separation than Reddit ELI5 and Wiki CSAI (general knowledge), consistent with Eq. 5's domain bias prediction.

particular human uses simpler words, while language models use complicated words with much less repetition. Fig. 10 shows the analysis results calculated per sample.

To facilitate reproducibility and further research in model-agnostic AI text detection, the dataset will be publicly released upon acceptance.

C.5. Domain Specialization Effects Summary

The domain-specific results validate our theoretical prediction (Eq. 5) that the magnitude of human-AI separation varies with domain specialization level. Specialized domains (Reuters, Finance, Medicine, Reddit_ELI5) with structured, technical vocabulary/way of writing exhibit the strongest separation. General-purpose domains (Creative Writing, Wiki CSAI) with diverse and unstructured vocabulary closer to $P_{\text{global}}(w)$ show diminished separation. When pooled across all domains, the clustering inequality tightens, but the Wall of Separation persists, confirming the robustness of letter distribution signatures for AI text detection.

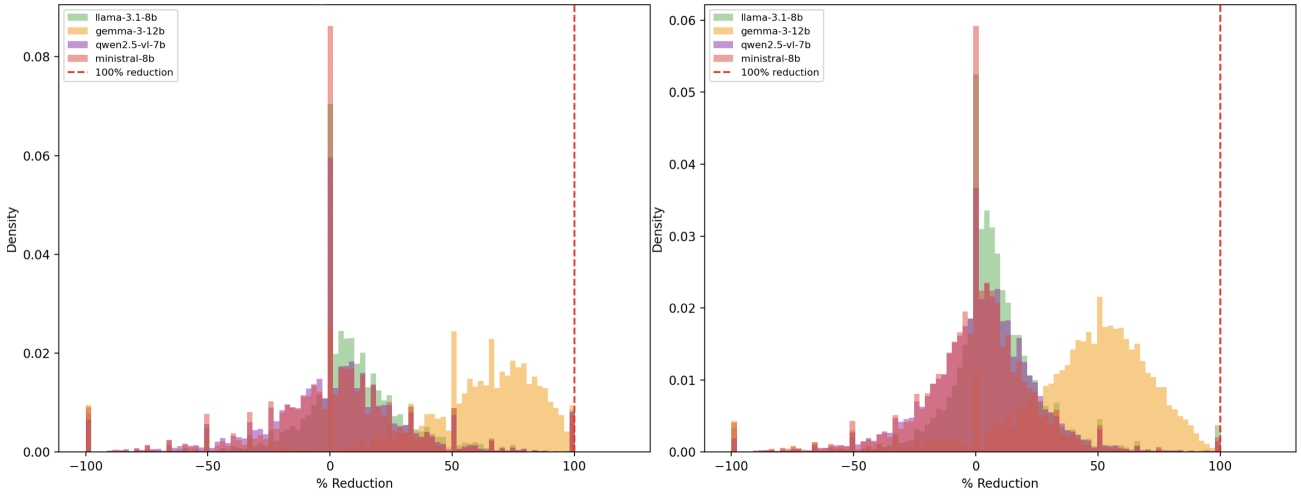


Figure 6. Distribution of percentage reduction in target letter frequency for the `ADV_X` (single-letter removal, left) and `ADV_XY` (two-letter removal, right) adversarial attacks, aggregated across all samples and domains in the MDTA dataset. The x-axis shows the percentage reduction in occurrences of the target letter(s) in the adversarial output relative to the original model response (temp $= 0.5$). A value of 100% (red dashed line) indicates complete removal of the target letter(s). Negative values indicate that the attack inadvertently *increased* the frequency of the target letter(s) in the output. `gemma-3-12b` achieves the highest reduction rates, while `ministral-8b` and `qwen2.5-vl-7b` frequently fail to reduce—or even add—target letters.

Table 6. N-gram counts across all domains in the MDTA dataset separated by source, calculated by concatenating all samples.

Source	Unigrams	Bigrams
Human	72,468	1,476,050
LLaMA-3.1-8B	50,484	1,377,721
Gemma-3-12B	43,059	1,299,946
Qwen2.5-VL-7B	45,167	1,369,694
Ministral-8B	42,824	1,078,912

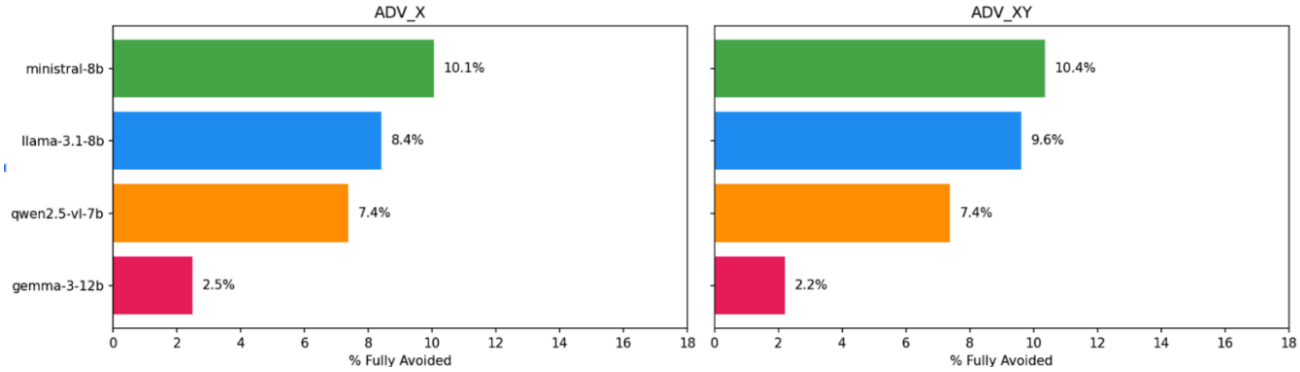


Figure 7. Percentage of samples where target letter(s) were fully absent from model outputs, averaged across all target letters, under single-letter (ADV_X) and two-letter (ADV_XY) constraints.

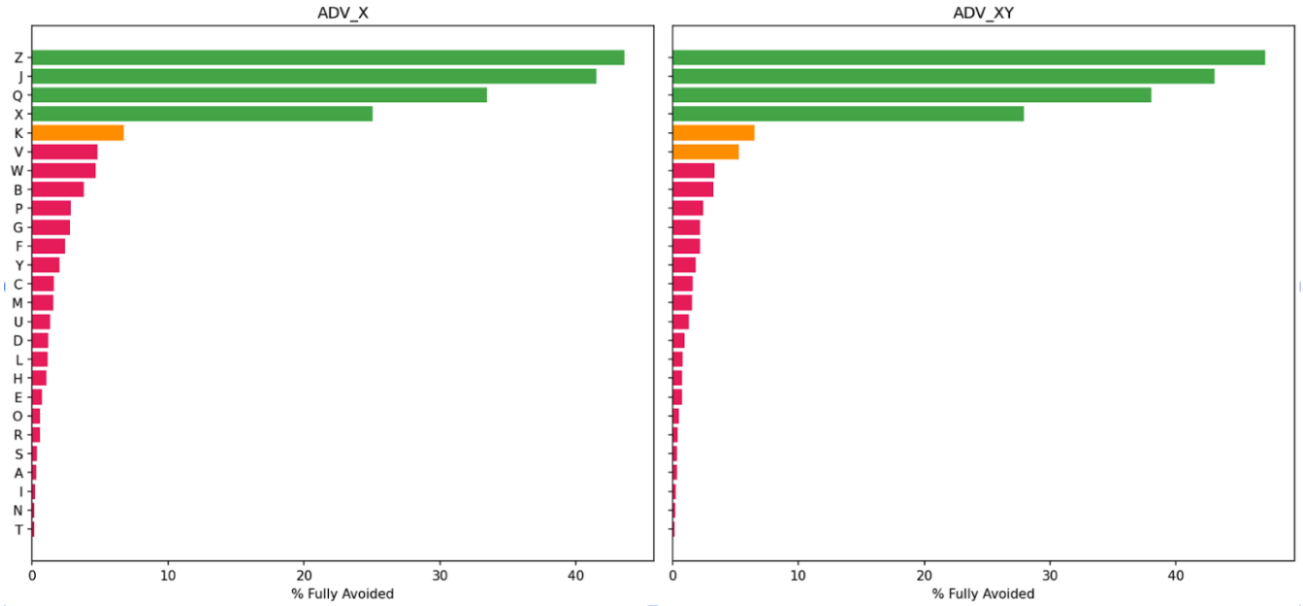


Figure 8. Per-letter avoidance success rate (100% removal) under ADV_X and ADV_XY, sorted in descending order.

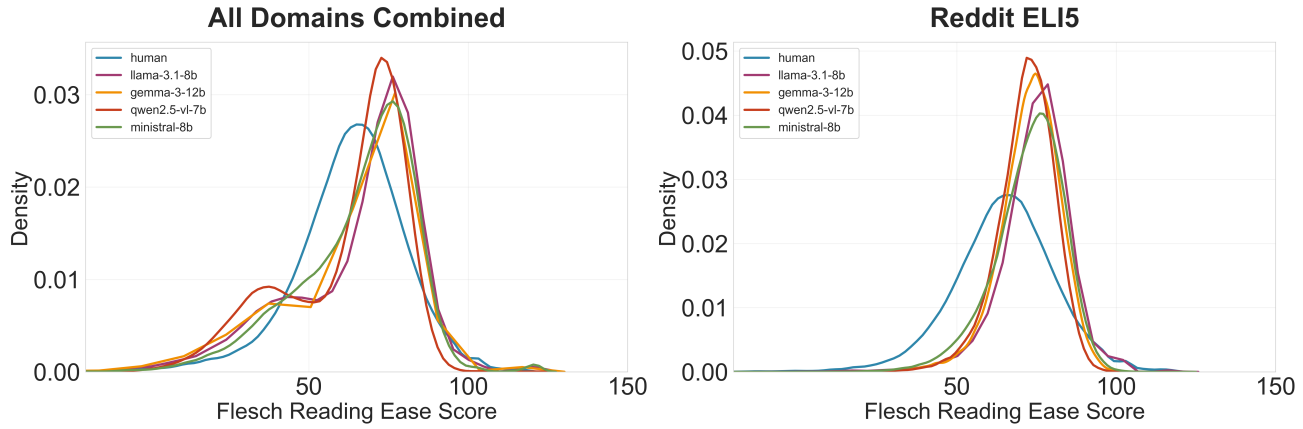


Figure 9. Instance-level readability analysis (FKGL Kernel Density Estimation (KDE)) across domains shows that human-written text is consistently simpler than model-generated text, with the gap most pronounced in the Reddit_ELI5 domain.

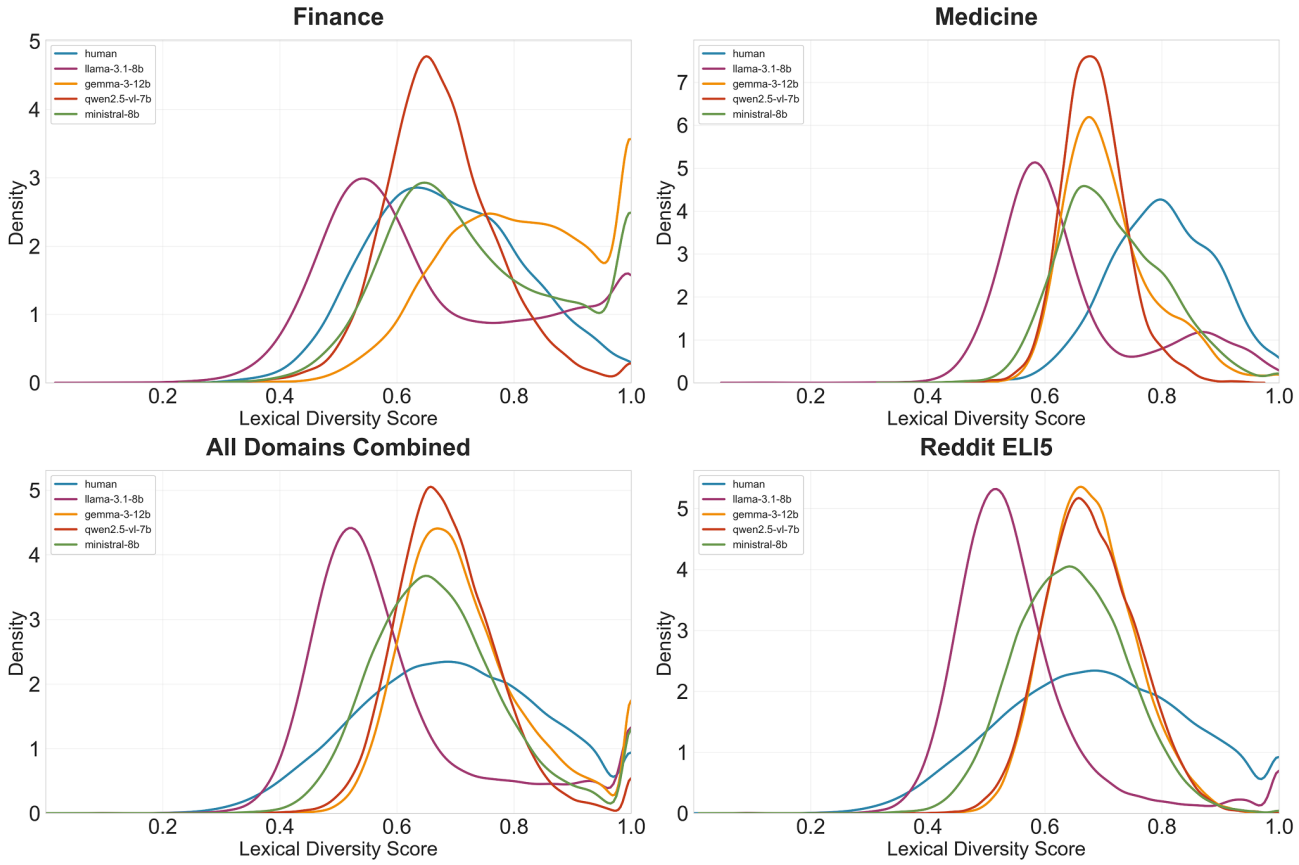


Figure 10. Instance-level lexical diversity (LDS) distributions across domains (Finance, Medicine, Reddit ELI5, and aggregated). Within specialized domains, human-written text exhibits LDS closer to those of AI-generated text, particularly in Finance and Medicine, while Reddit_ELI5 shows substantially higher LDS-variability between humans and AI.

Table 7. Detection performance on the Ghostbuster benchmark (Essay, Reuters, WP domains) with 100 balanced training samples. LD-X denotes augmentation of method X with the LD-Score via RBF-SVM fusion. Results are mean ± std over 5 runs (different seeds). **Bold** indicates best per column.

Method	Essay	Reuters	WP	Avg
AUROC				
DNA	**1.000**±0.000	0.999±0.003	0.984±0.009	0.994±0.005
Bino	**1.000**±0.001	0.995±0.005	0.961±0.012	0.985±0.008
FDGPT	0.977±0.013	0.962±0.016	0.865±0.026	0.935±0.019
LD-DNA	**1.000**±0.000	**1.000**±0.001	0.985±0.008	**0.995**±0.005
LD-Bino	**1.000**±0.000	0.999±0.002	0.967±0.008	0.989±0.005
LD-FDGPT	0.992±0.005	0.994±0.005	0.893±0.027	0.960±0.016
DNA+Bino	**1.000**±0.000	0.999±0.001	0.981±0.008	0.993±0.005
DNA+FDGPT	**1.000**±0.001	0.999±0.001	**0.986**±0.006	**0.995**±0.004
F1 Score				
DNA	0.991±0.006	0.978±0.012	0.945±0.021	0.971±0.014
Bino	0.981±0.018	0.952±0.030	0.895±0.032	0.943±0.027
FDGPT	0.918±0.024	0.893±0.025	0.780±0.044	0.864±0.032
LD-DNA	**0.996**±0.004	**0.996**±0.004	**0.949**±0.011	**0.980**±0.007
LD-Bino	0.990±0.005	0.988±0.006	0.918±0.021	0.965±0.013
LD-FDGPT	0.965±0.009	0.974±0.009	0.810±0.036	0.916±0.022
DNA+Bino	0.992±0.005	0.986±0.008	0.935±0.011	0.971±0.008
DNA+FDGPT	0.994±0.006	0.990±0.007	0.943±0.014	0.976±0.010

D. More Experiment Results

D.1. Ghostbuster Experiment Result

D.2. Results Based on Temperature.

The MDTA dataset also provides model responses generated at different temperature settings. We compare baseline and augmented methods across temperatures by running the analysis on the same human samples while replacing only the AI-generated inputs with responses from different temperatures. As a result, the False Positive Rate (FPR) remains constant. The results are summarized in Fig. 11. Our augmented approaches generally yield higher TPR with lower FPR. As expected, TPR typically decreases as the generation temperature increases.

As temperature increases, AI text exhibits greater word-level diversity, which degrades raw perplexity signals. However, this same variability moves AI text further from human letter distributions on average. Thus, while perplexity alone worsens, perplexity + lexical/LD signals benefit from temperature-induced diversity, increasing TPR and often lowering FPR by amplifying the global-vs-domain distribution gap.

D.3. Comparison with Stylometric Features

Recent work has explored rich stylometric feature sets for AI text detection. Przystalski et al. (2026) construct a suite of 196 linguistically-motivated features spanning lexical diversity, syntactic complexity, and punctuation patterns, identifying the most discriminative subset via SHAP analysis on a multiclass attribution task. We compare LD-score against their top-10

Table 8. F1 score comparison between LD-score and the top-10 stylometric features (ranked by multiclass SHAP importance from Przystalski et al. (2026)) trained on balanced subsets (per domain) of our benchmark and Ghostbuster dataset (125 human / 125 AI per domain, across 5 seeds). **Bold** denotes best per column.

Feature	Essay	Finance	Medicine	Open QA	Reddit ELI5	Reuters	Wiki CSAI	WP	Avg
Pooled RJSD	**0.84**	0.66	**0.87**	0.66	**0.78**	**0.93**	0.66	0.71	**0.76**
TTR Lemmas	0.71	0.67	0.76	0.67	0.66	0.74	0.67	0.77	0.71
Commas	0.67	0.67	0.74	0.66	0.71	0.70	0.67	**0.78**	0.70
Punctuation	0.66	**0.77**	0.79	0.67	0.69	0.67	0.66	0.67	0.70
Noun Phrase	0.66	0.72	0.76	0.66	0.72	0.67	0.66	0.72	0.70
Numerals	0.78	0.66	0.67	0.66	0.67	0.79	**0.68**	0.69	0.70
Words/Sent	0.67	0.66	0.67	**0.70**	0.70	0.78	0.66	0.72	0.70
Sent Diff	0.67	0.67	0.67	0.65	0.76	0.77	0.66	0.71	0.69
Dots	0.66	0.65	0.71	0.69	0.67	0.68	0.68	0.71	0.68
Fronting	0.66	0.67	0.68	0.67	0.67	0.71	0.67	0.69	0.68
L_FUNC_T	0.67	0.66	0.75	0.66	0.67	0.65	0.67	0.65	0.67

SHAP-ranked features, evaluated jointly across the MDTA and Ghostbuster datasets (125 human / 125 AI per domain).

Results. Table 8 reports F1 scores across all domains. LD-Score achieves the highest average F1 (0.76), outperforming all individual stylometric features. Gains are most pronounced in structured domains such as Reuters (0.93 vs. 0.78), Medicine (0.87 vs. 0.79), and Essay (0.84 vs. 0.71). This is consistent with our theoretical prediction that domain-specialized vocabularies amplify the letter-distribution signal.

Fragility of stylometric features. Despite the richness of the 196-feature suite, the highest-ranked features by SHAP importance are overwhelmingly simple surface-level statistics: punctuation frequency, comma rate, period count, words-per-sentence, and numeral density. These are precisely the features most vulnerable to adversarial manipulation-a model can trivially shift punctuation density or sentence length without altering semantic content. Shifting letter-level distributions, by contrast, requires coordinated vocabulary-level changes across the entire text. The dominance of surface features in the top-10 thus reflects a broader limitation of stylometry: in-distribution discriminability does not imply robustness under adaptive attack.

E. Adversarial Results

Table 9 reports AUROC under three adversarial conditions. The augmented variants (LD-DNA, LD-Bino) outperform their base counterparts in 29 of 30 domain-attack comparisons, with the single exception being `medicine` under attack (B), where the difference is negligible (-0.001 and -0.002 respectively). Gains are largest under attack (B) (single letter removal), where Binoculars improves by up to $+0.029$ on `reddit_eli5` and DNA improves by $+0.015$ on `finance`, suggesting the character-level LD-Score signal is particularly complementary when token distributions are locally perturbed. Attack (A) (paraphrase) yields the highest absolute AUROC values across all methods, with averages of 0.950 and 0.944 for DNA+LD and Bino+LD respectively, indicating that paraphrasing alone does not substantially degrade detection. Attack (C) (dual letter removal) is the hardest condition, producing the lowest overall AUROC, yet augmentation still improves both detectors consistently, with average gains of $+0.006$ and $+0.012$ for DNA and Binoculars. Across all attacks, `open_qa` remains the most challenging domain, while `medicine` and `finance` are consistently the easiest.

Table 10 reports F1 scores under the same three conditions. Unlike AUROC, augmentation yields mixed results: under attack (C) (dual letter removal), both LD-DNA and LD-Bino degrade on `open_qa` (-0.091 and -0.082 respectively), and the overall averages are comparable or slightly worse than the base detectors. Attacks (A) and (B) tell a more favorable story, with consistent gains on `medicine`, `reddit_eli5`, and `wiki_csai`, and average improvements of $+0.008$ for both LD-DNA and LD-Bino under paraphrasing. As with AUROC, `open_qa` remains the most adversarially vulnerable domain across all attacks and methods.

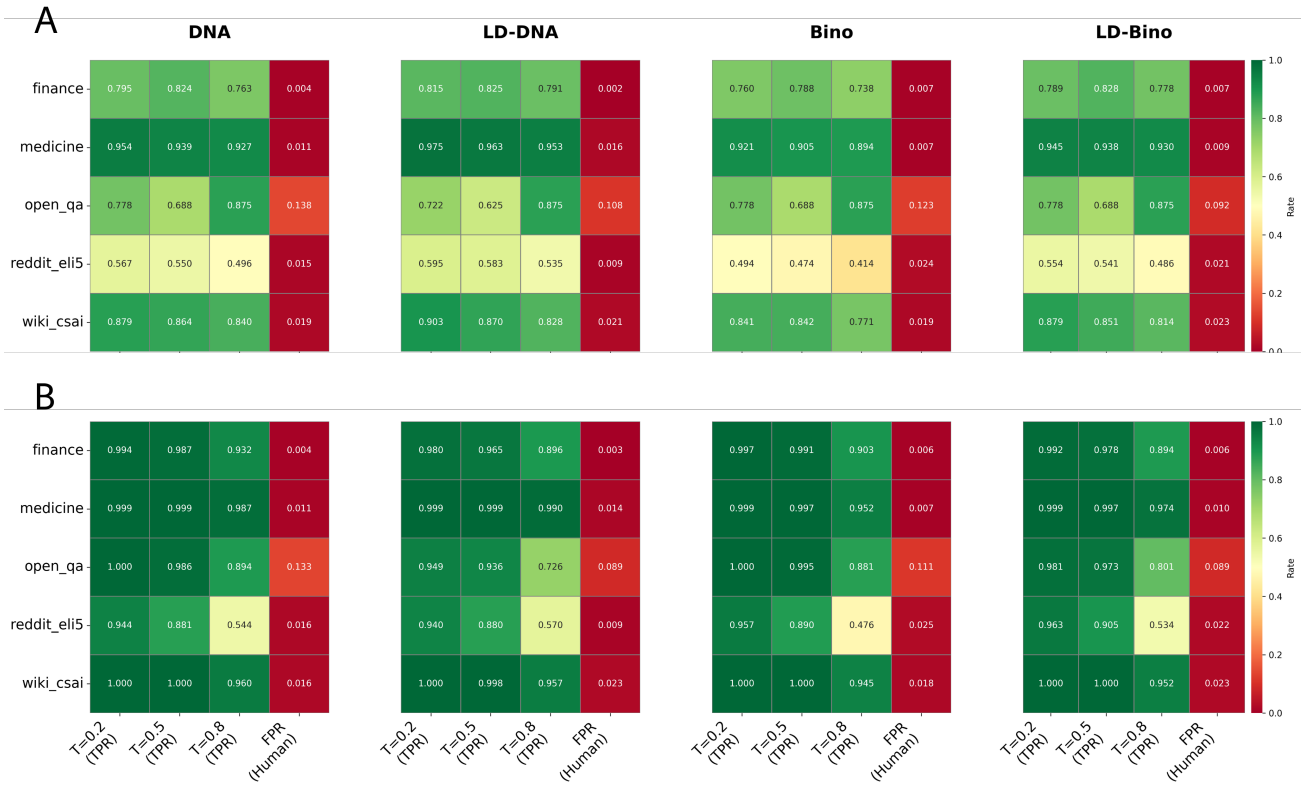


Figure 11. Temperature-dependent detection performance across domains for Gemma-3-12B **(A)** and Qwen2.5-VL-7B **(B)** models. Heatmaps show True Positive Rate (TPR) at different generation temperatures (T=0.2, 0.5, 0.8) and False Positive Rate (FPR) on human text across five domains and four detection methods. All methods use globally optimized thresholds. Increasing the temperature typically worsens the TPR. Augmenting with our approach mostly increase the TPR and reduces FPR, although there are exceptions.

Table 9. AUROC by Domain and Attack Type. (A): paraphrase attack, (B): single letter removal, (C): dual letter removal. **Bold** entries indicate which of the augmented or unaugmented variant performed better.

Attack	Dataset	DNA	LD-DNA	Δ	Bino	LD-Bino	Δ
(A)	finance	$0.975_{\pm 0.009}$	$\mathbf{0.979}_{\pm 0.011}$	+0.004	$0.968_{\pm 0.013}$	$\mathbf{0.978}_{\pm 0.007}$	+0.010
	medicine	$0.995_{\pm 0.002}$	$\mathbf{0.996}_{\pm 0.003}$	+0.001	$0.994_{\pm 0.004}$	$\mathbf{0.995}_{\pm 0.006}$	+0.001
	open_qa	$0.834_{\pm 0.012}$	$\mathbf{0.845}_{\pm 0.025}$	+0.011	$0.808_{\pm 0.011}$	$\mathbf{0.828}_{\pm 0.028}$	+0.020
	reddit_eli5	$0.938_{\pm 0.008}$	$\mathbf{0.948}_{\pm 0.014}$	+0.010	$0.922_{\pm 0.009}$	$\mathbf{0.937}_{\pm 0.013}$	+0.015
	wiki_csai	$0.982_{\pm 0.007}$	$\mathbf{0.984}_{\pm 0.004}$	+0.002	$0.979_{\pm 0.006}$	$\mathbf{0.983}_{\pm 0.004}$	+0.004
	avg	$0.945_{\pm 0.008}$	$\mathbf{0.950}_{\pm 0.011}$	+0.005	$0.934_{\pm 0.009}$	$\mathbf{0.944}_{\pm 0.012}$	+0.010
(B)	finance	$0.946_{\pm 0.005}$	$\mathbf{0.961}_{\pm 0.005}$	+0.015	$0.916_{\pm 0.014}$	$\mathbf{0.938}_{\pm 0.013}$	+0.022
	medicine	$\mathbf{0.980}_{\pm 0.005}$	$0.979_{\pm 0.003}$	−0.001	$\mathbf{0.964}_{\pm 0.007}$	$0.962_{\pm 0.006}$	−0.002
	open_qa	$0.756_{\pm 0.042}$	$\mathbf{0.761}_{\pm 0.026}$	+0.005	$0.719_{\pm 0.042}$	$\mathbf{0.727}_{\pm 0.032}$	+0.008
	reddit_eli5	$0.877_{\pm 0.020}$	$\mathbf{0.886}_{\pm 0.024}$	+0.009	$0.828_{\pm 0.027}$	$\mathbf{0.857}_{\pm 0.022}$	+0.029
	wiki_csai	$0.964_{\pm 0.013}$	$\mathbf{0.968}_{\pm 0.010}$	+0.004	$0.943_{\pm 0.017}$	$\mathbf{0.954}_{\pm 0.013}$	+0.011
	avg	$0.905_{\pm 0.017}$	$\mathbf{0.911}_{\pm 0.014}$	+0.006	$0.874_{\pm 0.021}$	$\mathbf{0.888}_{\pm 0.017}$	+0.014
(C)	finance	$0.938_{\pm 0.014}$	$\mathbf{0.936}_{\pm 0.013}$	−0.002	$0.898_{\pm 0.016}$	$\mathbf{0.902}_{\pm 0.018}$	+0.004
	medicine	$0.959_{\pm 0.011}$	$\mathbf{0.961}_{\pm 0.010}$	+0.002	$0.943_{\pm 0.009}$	$\mathbf{0.946}_{\pm 0.012}$	+0.003
	open_qa	$0.712_{\pm 0.020}$	$\mathbf{0.727}_{\pm 0.034}$	+0.015	$0.674_{\pm 0.023}$	$\mathbf{0.691}_{\pm 0.033}$	+0.017
	reddit_eli5	$0.853_{\pm 0.032}$	$\mathbf{0.863}_{\pm 0.028}$	+0.010	$0.798_{\pm 0.037}$	$\mathbf{0.819}_{\pm 0.017}$	+0.021
	wiki_csai	$0.944_{\pm 0.017}$	$\mathbf{0.946}_{\pm 0.012}$	+0.002	$0.914_{\pm 0.022}$	$\mathbf{0.928}_{\pm 0.021}$	+0.014
	avg	$0.881_{\pm 0.019}$	$\mathbf{0.887}_{\pm 0.019}$	+0.006	$0.845_{\pm 0.021}$	$\mathbf{0.857}_{\pm 0.020}$	+0.012
Avg		$0.910_{\pm 0.014}$	$\mathbf{0.916}_{\pm 0.015}$	+0.006	$0.885_{\pm 0.017}$	$\mathbf{0.896}_{\pm 0.016}$	+0.011

Table 10. F1 Score by Domain and Attack Type. (A): paraphrase attack, (B): single letter removal, (C): dual letter removal. **Bold** entries indicate which of the augmented or unaugmented variant performed better.

Attack	Dataset	DNA	LD-DNA	Δ	Bino	LD-Bino	Δ
(A)	finance	$\mathbf{0.946}_{\pm 0.012}$	$0.946_{\pm 0.011}$	$+0.000$	$0.929_{\pm 0.016}$	$\mathbf{0.933}_{\pm 0.012}$	$+0.004$
	medicine	$0.973_{\pm 0.009}$	$\mathbf{0.981}_{\pm 0.008}$	$+0.008$	$0.968_{\pm 0.016}$	$\mathbf{0.978}_{\pm 0.011}$	$+0.010$
	open_qa	$0.778_{\pm 0.014}$	$\mathbf{0.785}_{\pm 0.020}$	$+0.007$	$0.748_{\pm 0.019}$	$\mathbf{0.750}_{\pm 0.026}$	$+0.002$
	reddit_eli5	$0.869_{\pm 0.010}$	$\mathbf{0.886}_{\pm 0.016}$	$+0.017$	$0.841_{\pm 0.015}$	$\mathbf{0.873}_{\pm 0.010}$	$+0.032$
	wiki_csai	$0.936_{\pm 0.021}$	$\mathbf{0.941}_{\pm 0.018}$	$+0.005$	$\mathbf{0.937}_{\pm 0.019}$	$0.933_{\pm 0.011}$	-0.004
	avg	$0.900_{\pm 0.013}$	$\mathbf{0.908}_{\pm 0.015}$	$+0.008$	$0.885_{\pm 0.017}$	$\mathbf{0.893}_{\pm 0.014}$	$+0.008$
(B)	finance	$0.905_{\pm 0.014}$	$\mathbf{0.909}_{\pm 0.019}$	$+0.004$	$0.854_{\pm 0.030}$	$\mathbf{0.864}_{\pm 0.028}$	$+0.010$
	medicine	$\mathbf{0.924}_{\pm 0.017}$	$0.923_{\pm 0.008}$	-0.001	$0.884_{\pm 0.023}$	$\mathbf{0.891}_{\pm 0.010}$	$+0.007$
	open_qa	$\mathbf{0.720}_{\pm 0.049}$	$0.692_{\pm 0.020}$	-0.028	$\mathbf{0.694}_{\pm 0.040}$	$0.656_{\pm 0.030}$	-0.038
	reddit_eli5	$0.781_{\pm 0.023}$	$\mathbf{0.806}_{\pm 0.014}$	$+0.025$	$0.740_{\pm 0.033}$	$\mathbf{0.755}_{\pm 0.025}$	$+0.015$
	wiki_csai	$0.907_{\pm 0.023}$	$\mathbf{0.915}_{\pm 0.025}$	$+0.008$	$0.878_{\pm 0.016}$	$\mathbf{0.896}_{\pm 0.018}$	$+0.018$
	avg	$0.847_{\pm 0.025}$	$\mathbf{0.849}_{\pm 0.017}$	$+0.002$	$0.810_{\pm 0.028}$	$\mathbf{0.812}_{\pm 0.022}$	$+0.002$
(C)	finance	$\mathbf{0.880}_{\pm 0.025}$	$0.879_{\pm 0.029}$	-0.001	$0.825_{\pm 0.023}$	$\mathbf{0.841}_{\pm 0.025}$	$+0.016$
	medicine	$0.886_{\pm 0.021}$	$\mathbf{0.898}_{\pm 0.013}$	$+0.012$	$0.858_{\pm 0.016}$	$\mathbf{0.876}_{\pm 0.023}$	$+0.018$
	open_qa	$\mathbf{0.692}_{\pm 0.017}$	$0.601_{\pm 0.035}$	-0.091	$\mathbf{0.667}_{\pm 0.011}$	$0.585_{\pm 0.036}$	-0.082
	reddit_eli5	$0.774_{\pm 0.023}$	$\mathbf{0.784}_{\pm 0.015}$	$+0.010$	$\mathbf{0.718}_{\pm 0.027}$	$0.718_{\pm 0.018}$	$+0.000$
	wiki_csai	$0.875_{\pm 0.014}$	$\mathbf{0.888}_{\pm 0.024}$	$+0.013$	$0.837_{\pm 0.027}$	$\mathbf{0.877}_{\pm 0.021}$	$+0.040$
	avg	$\mathbf{0.821}_{\pm 0.020}$	$0.810_{\pm 0.023}$	-0.011	$\mathbf{0.781}_{\pm 0.021}$	$0.779_{\pm 0.025}$	-0.002
Avg		$\mathbf{0.856}_{\pm 0.019}$	$0.856_{\pm 0.018}$	$+0.000$	$0.825_{\pm 0.022}$	$\mathbf{0.828}_{\pm 0.020}$	$+0.003$