

TCD: Empowering Zero-Shot Detection of LLM-Generated Text using Token Contextual Dependency

Shihan Ma, Quan Wang*

MOE Key Laboratory of Trustworthy Distributed Computing and Service,
Beijing University of Posts and Telecommunications, Beijing, China
{mashihan, wangquan}@bupt.edu.cn

Abstract

Zero-shot detectors for LLM-generated text aggregate token-level statistics uniformly, implicitly assuming that every position contributes equally to detection. We challenge this assumption and show that detection signal is highly uneven across token positions: positions of high contextual dependency, where the surrogate model relies on concentrated context, carry far more discriminative power than others. Using attention entropy to quantify this dependency, we empirically validate that aggregating only low-entropy tokens yields clearly separated human and LLM score distributions, whereas high-entropy tokens do not. Based on this finding, we propose Token Contextual Dependency (TCD), a plug-and-play framework that reweights token contributions in existing zero-shot detectors via attention entropy-based weights computed from dynamically selected layers. Experiments on state-of-the-art detectors across multiple benchmarks demonstrate consistent and meaningful improvements in detection performance. Our code is available at: <https://github.com/RecursionM/TCD>.

1 Introduction

The rapid advancement of large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) has transformed them into indispensable tools for content creation, enabling the generation of high-quality text at scale across domains ranging from journalism and education to scientific writing (Demszky et al., 2023; Diakopoulos et al., 2024; Liang et al., 2024). However, this capability has also raised serious concerns regarding the misuse of LLM-generated content, including the spread of misinformation, academic dishonesty, and plagiarism (Baarir and Djeflal, 2021; Pan et al., 2023; Eke, 2023; Pudasaini et al., 2025; Ibrahim et al., 2023; Friederich and Symons, 2023). Con-

sequently, developing accurate and reliable techniques for detecting LLM-generated text has therefore become an urgent and critical research challenge.

Existing detection methods are broadly categorized into supervised and zero-shot approaches. Supervised methods train classifiers on labeled datasets of human-written and LLM-generated text, but are prone to overfitting and often fail to generalize across different domains or LLM architectures (Guo et al., 2024; Verma et al., 2024; Hu et al., 2023). Zero-shot methods, in contrast, require no training data and instead leverage statistical differences between the two text types directly from surrogate models, offering substantially better generalization (Mitchell et al., 2023; Bao et al., 2024). Consequently, zero-shot detection has become the primary focus of current research, operating by using a surrogate model to compute next-token prediction probabilities for each position given its preceding context, and deriving token-level probability statistics to determine whether the text is LLM-generated (Gehrmann et al., 2019; Su et al., 2023; Hans et al., 2024; Zhu et al., 2026a).

However, a critical yet overlooked assumption underlies these methods: they aggregate token-level statistics uniformly, implicitly treating every position as equally informative for detection. We challenge this assumption. Our central hypothesis is that detection signal is highly uneven across token positions: positions where the surrogate model’s prediction of the current token relies on a concentrated subset of preceding context, i.e., positions of high *contextual dependency*, carry far more discriminative power than positions where the prediction is spread diffusely across context. Intuitively, concentrated dependency suggests that the model is applying a formulaic, pretraining-familiar contextual pattern; LLM-generated text tends to follow such collective patterns, whereas human writing is more idiosyncratic. Detector scores

*Corresponding author: Quan Wang.

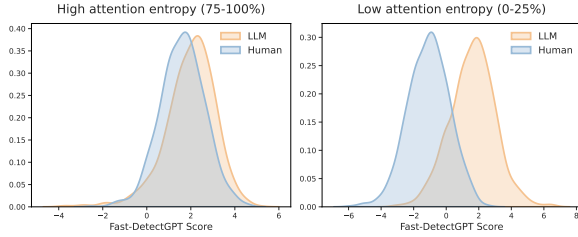


Figure 1: Empirical validation of the contextual-dependency hypothesis using Fast-DetectGPT (Bao et al., 2024) on 2000 balanced examples sampled from RealDet (Zhu et al., 2025). **Left:** scores computed with only the highest-attention-entropy tokens (75–100%). High-entropy (low-dependency) tokens yield largely overlapping human vs. LLM distributions. **Right:** scores computed with only the lowest-entropy tokens (0–25%). Low-entropy (high-dependency) tokens yield clearly separated human vs. LLM distributions.

should therefore diverge more between LLM and human text at high-dependency positions, while low-dependency positions should contribute little to discrimination.

To operationalize contextual dependency and test this hypothesis, We quantify how concentrated each token’s reliance on context is using the *attention entropy* of the surrogate model. Surrogate models in zero-shot detection are typically Transformer-based (Vaswani et al., 2017); in this setting, a sharply peaked attention distribution means the next-token prediction is anchored to few context tokens (low entropy $\Rightarrow$ high dependency), whereas a flat distribution indicates diffuse reliance (high entropy $\Rightarrow$ low dependency). As shown in Figure 1, human and LLM score distributions largely overlap when only high-entropy tokens are used (*left*), but separate clearly when only low-entropy tokens are used (*right*), confirming that high-dependency positions concentrate the detection signal.

Building on this evidence, we propose **Token Contextual Dependency (TCD)**, a plug-and-play framework that reweights token contributions in existing zero-shot detectors. Specifically, we measure per-token attention entropy at each layer and introduce *dynamic layer selection* (DLS), as different layers capture qualitatively different information. DLS retains layers whose mean entropy falls below the global average, as they contain more focused-dependency tokens and richer signal, and layers in the latter half of the network, where higher-order semantic features better separate LLM-generated from human-written text (Jiao et al., 2025). We aggregate attention entropy over these selected layers

to obtain a TCD score for each position; lower entropy indicates stronger contextual dependency and receives higher weight via softmax. The resulting weights replace uniform averaging with a weighted sum over token-level detector statistics.

To evaluate the effectiveness of our approach, we apply it to three state-of-the-art zero-shot detectors (Bao et al., 2024; Hans et al., 2024; Zhu et al., 2026a) and conduct extensive experiments on four widely used public benchmarks (Zhu et al., 2025; Dugan et al., 2024; Wang et al., 2024; Bao et al., 2025). Experimental results demonstrate consistent and meaningful improvements across all three detectors, validating TCD as a plug-and-play enhancement to existing zero-shot detection. Our main contributions are as follows:

- We reveal that detection signal is highly uneven across token positions: positions of high contextual dependency carry far more discriminative power than positions of low contextual dependency.
- We propose TCD, a plug-and-play token weighting framework that operationalizes contextual dependency via attention entropy and a dynamic layer selection strategy.
- We apply TCD into state-of-the-art zero-shot detectors and empirically demonstrate the effectiveness of our approach across diverse public benchmarks.

2 Related Work

2.1 LLM-Generated Text Detection

Identifying LLM-generated text is usually cast as distinguishing human writing from LLM output (Tang et al., 2024). Supervised detectors fine-tune on labeled corpora (Guo et al., 2024; Verma et al., 2024; Hu et al., 2023) but often degrade on unseen domains, generators, or edited text. Zero-shot methods have therefore received growing attention (Mitchell et al., 2023; Bao et al., 2024). These approaches score a passage with surrogate language models using token-level statistics such as likelihood, log-rank, and probability curvature (Mitchell et al., 2023; Su et al., 2023; Bao et al., 2024), and recent multi-model variants further compare discrepancies between paired models (Hans et al., 2024; Zhu et al., 2026a). Representative systems include DetectGPT (Mitchell et al.,

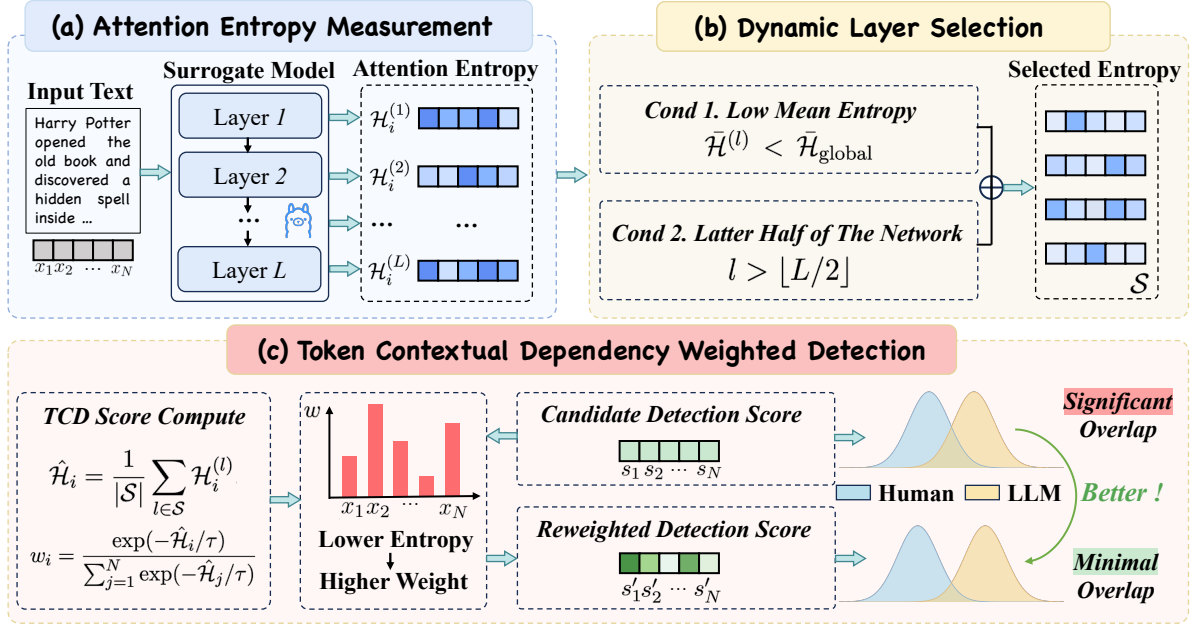


Figure 2: Overview of the TCD framework. (a) **Attention entropy measurement**: input tokens are passed through a surrogate model to compute per-token attention entropy $\mathcal{H}_i^{(l)}$ at each layer. (b) **Dynamic layer selection (DLS)**: layers are retained only if their mean entropy is below the global average and if they lie in the latter half of the network, yielding a selected layer set $\mathcal{S}$. (c) **Token contextual dependency weighted detection**: entropy over $\mathcal{S}$ is aggregated into a TCD score per token, converted to softmax weights (lower entropy $\Rightarrow$ higher weight), and used to reweight token-level detector statistics, yielding better separation between human and LLM-generated text than uniform aggregation.

2023), Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), and DNA-DetectLLM (Zhu et al., 2026a). Although their scoring functions differ, most of them combine token-level values through uniform averaging, which is the assumption challenged in this paper.

2.2 Attention-Based LLM Output Analysis

Transformer self-attention maps each token to its context (Vaswani et al., 2017). In LLM output analysis, attention structure has been linked to hallucination: Chuang et al. (Chuang et al., 2024) show that contextual hallucinations correlate with reduced attention to source context relative to generated tokens, and Wu et al. (Wu et al., 2025) identify retrieval attention heads whose activation determines whether generations remain faithful to the input or hallucinate. These findings indicate that per-token attention carries interpretable signal about output reliability, motivating TCD to use attention entropy for position-aware reweighting in zero-shot detection.

3 Methodology

Figure 2 illustrates TCD, a plug-and-play framework that reweights token contributions in existing

zero-shot detectors according to contextual dependency. TCD proceeds in three stages aligned with the figure: (a) measuring per-token attention entropy via a surrogate model (Section 3.2); (b) selecting informative layers through Dynamic Layer Selection (DLS, Section 3.3); and (c) aggregating entropy into token weights and integrating them into zero-shot detection (Section 3.4). We first formalize the uniform-aggregation baseline that TCD replaces in Section 3.1.

3.1 Preliminary

Zero-shot LLM-generated text detection is formulated as a binary classification problem. Given a candidate passage $x = \{x_1, x_2, \dots, x_N\}$ of length N , the goal is to determine whether x is human-written or LLM-generated without any labeled training data. Existing methods employ one or more language models to compute a token-level statistic s_i at each position and aggregate these statistics uniformly into a detection score:

$$S(x) = f\left(\frac{1}{N} \sum_{i=1}^N s_i\right), \quad (1)$$

where $f(\cdot)$ is a detector-specific transformation. We denote this uniformly aggregated score as the

candidate detection score. As argued in Section 1, it implicitly treats every position as equally informative, overlooking the uneven distribution of detection signal across token positions.

3.2 Attention Entropy Measurement

As shown in Figure 2(a), we quantify contextual dependency by passing x through a surrogate Transformer model with L layers and computing a per-token attention entropy vector $\mathcal{H}_i^{(l)}$ at each layer l . A sharply peaked attention distribution indicates that the prediction at position i is anchored to a concentrated subset of preceding context (low entropy $\Rightarrow$ high dependency), whereas a flat distribution indicates diffuse reliance (high entropy $\Rightarrow$ low dependency). Positions of high contextual dependency are therefore expected to carry richer discriminative signal, as empirically validated in Figure 1.

Formally, let $\mathbf{A}^{(l)} \in \mathbb{R}^{H \times N \times N}$ denote the attention weight matrix at layer l , where H is the number of attention heads. The attention entropy of token i at layer l is:

$$\mathcal{H}_i^{(l)} = -\frac{1}{H} \sum_{h=1}^H \sum_{j=1}^N A_{h,i,j}^{(l)} \log A_{h,i,j}^{(l)}, \quad (2)$$

where $A_{h,i,j}^{(l)}$ is the attention weight from token i to token j at head h in layer l . Lower values of $\mathcal{H}_i^{(l)}$ correspond to more focused attention and stronger contextual dependency at position i .

3.3 Dynamic Layer Selection

Figure 2(b) summarizes DLS. Although attention entropy can be computed at all L layers, different layers capture qualitatively different information, and not all layers contribute equally to discriminating LLM-generated from human-written text. DLS therefore retains only layers satisfying *both* of the following conditions, yielding a selected layer set $\mathcal{S}$.

Condition 1: Low mean entropy. For each layer l , we compute its mean attention entropy across all token positions:

$$\bar{\mathcal{H}}^{(l)} = \frac{1}{N} \sum_{i=1}^N \mathcal{H}_i^{(l)}, \quad (3)$$

and the global mean across all layers as $\bar{\mathcal{H}}_{\text{global}} = \frac{1}{L} \sum_{l=1}^L \bar{\mathcal{H}}^{(l)}$. A layer is included only if $\bar{\mathcal{H}}^{(l)} < \bar{\mathcal{H}}_{\text{global}}$, ensuring that retained layers contain more focused-dependency tokens and richer signal.

Condition 2: Latter half of the network. A layer is included only if $l > \lfloor L/2 \rfloor$. Earlier layers primarily capture shallow lexical and syntactic patterns on which LLM-generated and human-written texts are difficult to distinguish, whereas deeper layers encode higher-order semantic features that exhibit more pronounced divergence between the two text types.

Both conditions follow from properties shared by Transformer LLMs rather than from the attention pattern of a particular checkpoint, and $\mathcal{S}$ is re-derived for each input, so DLS transfers across surrogate models and domains (Appendix B.5).

3.4 Token Contextual Dependency Weighted Detection

Figure 2(c) shows how selected entropy is converted into detection weights. Let $\mathcal{S}$ denote the layer set selected by DLS. We aggregate per-token attention entropy over $\mathcal{S}$ to obtain the Token Contextual Dependency (TCD) score of each position:

$$\hat{\mathcal{H}}_i = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \mathcal{H}_i^{(l)}, \quad (4)$$

where a lower value of $\hat{\mathcal{H}}_i$ indicates stronger contextual dependency and greater expected discriminative value for detection. We convert $\hat{\mathcal{H}}_i$ into a normalized token weight via temperature-scaled softmax:

$$w_i = \frac{\exp(-\hat{\mathcal{H}}_i/\tau)}{\sum_{j=1}^N \exp(-\hat{\mathcal{H}}_j/\tau)}, \quad (5)$$

where $\tau > 0$ controls the sharpness of the weight distribution. Lower entropy yields higher weight, as annotated in Figure 2(c).

Given token-level detector statistics s_i from an existing zero-shot method, the uniform aggregation in Eq. (1) treats all positions equally. TCD instead assigns a reweighted statistic $s'_i = w_i \cdot s_i$ to each position and computes a *reweighted detection score*:

$$S'(x) = f\left(\sum_{i=1}^N s'_i\right) = f\left(\sum_{i=1}^N w_i \cdot s_i\right), \quad (6)$$

emphasizing high-dependency positions while suppressing low-signal positions. Compared with the candidate score $S(x)$ in Eq. (1), $S'(x)$ yields sharper separation between human-written and LLM-generated text. TCD is entirely plug-and-play: it requires only a single forward pass through

the surrogate model with attention outputs enabled, and integrates seamlessly into any zero-shot detector that aggregates token-level statistics.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TCD on three widely used public benchmarks for LLM-generated text detection. **RealDet** is a large-scale bilingual corpus spanning diverse domains and source LLMs, designed to reflect realistic deployment conditions (Zhu et al., 2025). **RAID** is a challenging robustness benchmark covering multiple domains, source models, decoding strategies, and adversarial perturbations (Dugan et al., 2024). **M4** is a multi-generator, multi-domain, and multi-lingual corpus for machine-generated text detection (Wang et al., 2024). For each benchmark, we use a balanced subset of 1,000 human-written and 1,000 LLM-generated passages. Beyond the main evaluation, we conduct additional analyses on: (i) the multi-domain partition of RAID, where each domain contains text from multiple source LLMs; (ii) the HART benchmark for hybrid text detection across three difficulty levels (Bao et al., 2025); (iii) the adversarial attack subset of RAID; and (iv) length-truncated variants of RAID.

Baselines. TCD is designed as a plug-and-play enhancement and is applied on top of three strong zero-shot detectors: Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), and DNA-DetectLLM (Zhu et al., 2026a). We also report results for four representative probability-based baselines, namely Likelihood, LogRank, and Entropy from GLTR (Gehrmann et al., 2019) and LRR (Su et al., 2023), to contextualize overall detection performance. All TCD-augmented variants share the same base detector configuration and differ only in the entropy-based token reweighting introduced in Section 3.

Metrics. Following standard practice in zero-shot detection (Bao et al., 2024; Mitchell et al., 2023), we report the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPR). Both metrics operate over continuous detection scores and do not require a fixed decision threshold, which is especially important in zero-shot settings where score distributions may shift across domains or generators. We report percentages throughout for

readability.

Implementation Details. Unless otherwise stated, we use Llama-3.2-1B-Instruct (Grattafiori et al., 2024) as the surrogate model for attention entropy computation and scoring, paired with Llama-3.2-1B where a base/reference model is required by the underlying detector, chosen for its small size and correspondingly low detection latency. All experiments are conducted on NVIDIA A40 GPUs (48GB each), with a maximum input length of 512 tokens and a batch size of 8. The softmax temperature in Section 3.4 is set to $\tau = 0.7$ by default. All base detectors retain their original scoring logic; TCD only enabled attention output during the forward propagation through the surrogate model once.

4.2 Main Results

Table 1 summarizes detection performance on RealDet, RAID, and M4. Across all three benchmarks, applying TCD yields consistent improvements over every base detector. On average, TCD improves Fast-DetectGPT by 3.13% AUROC and 1.37% AUPR, Binoculars by 2.32% AUROC and 0.91% AUPR, and DNA-DetectLLM by 2.23% AUROC and 1.03% AUPR. The largest absolute AUROC gain appears on RAID (+4.45% for Fast-DetectGPT), which is also the most challenging benchmark among the three, suggesting that emphasizing high-dependency tokens is particularly helpful when human and LLM score distributions overlap more heavily. Importantly, TCD never degrades any base method: even on RealDet and M4, where the baselines are already strong, the reweighted scores further separate human-written from LLM-generated text. These results confirm that contextual-dependency weighting provides a general and complementary signal beyond uniform token aggregation.

4.3 Analysis

Multi Domains and Multi Source Models. To assess cross-domain and cross-generator generalization, we evaluate on the multi-domain partition of RAID, covering eight genres: abstracts, books, news, poetry, recipes, Reddit posts, reviews, and Wikipedia (Figure 3). TCD improves all three base detectors in every domain, with particularly notable gains on poetry (+5.46% AUROC for Fast-DetectGPT+TCD) and books (+3.28%). Smaller but consistent improvements on news and wiki indi-

Methods	RealDet		RAID		M4		Avg.	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
<i>Zero-Shot Detectors</i>								
Likelihood	88.09	91.75	77.51	83.93	88.81	89.51	84.81	88.40
LogRank	88.91	92.25	78.46	83.95	89.91	90.76	85.76	88.99
Entropy	87.35	90.52	75.20	79.41	86.72	87.17	83.09	85.70
LRR	88.87	91.31	80.01	80.64	88.74	90.72	85.87	87.56
Fast-DetectGPT	90.54	93.55	82.80	89.75	89.54	92.25	87.63	91.85
+TCD (Ours)	93.56	95.10	87.25	91.44	91.48	93.12	90.76	93.22
(Diff)	3.02	1.55	4.45	1.69	1.94	0.87	3.13	1.37
Binoculars	91.60	94.15	83.10	89.94	89.12	91.28	87.94	91.79
+TCD (Ours)	93.02	94.51	86.88	91.19	90.90	92.39	90.26	92.70
(Diff)	1.42	0.36	3.78	1.25	1.78	1.11	2.32	0.91
DNA-DetectLLM	91.86	94.22	82.16	88.90	90.25	91.84	88.09	91.65
+TCD (Ours)	93.36	94.81	85.24	89.51	92.37	93.70	90.32	92.68
(Diff)	1.50	0.59	3.08	0.61	2.12	1.86	2.23	1.03

Table 1: Comparison of detection performance (AUROC and AUPR, %) on public benchmark datasets. “+TCD” denotes our plug-and-play token weighting framework applied to the base method. Best results are **bolded**.

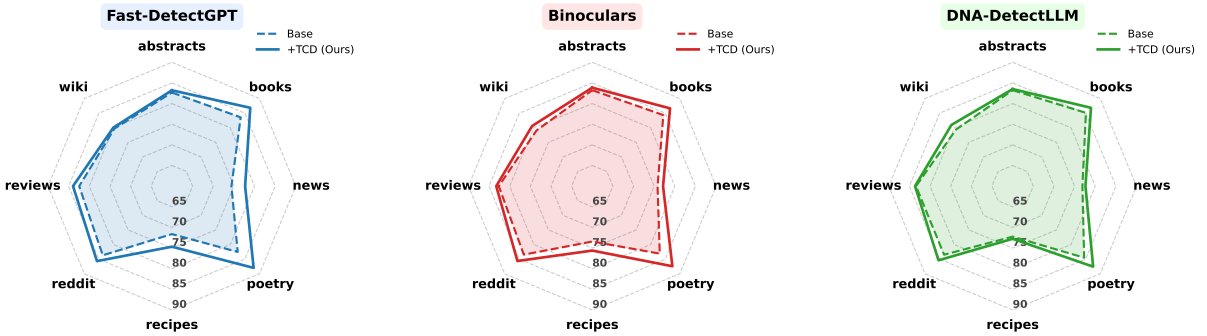


Figure 3: Multi domains radar chart. The results show the AUROC of three baseline models (Fast-DetectGPT, Binoculars, and DNA-DetectLLM) in eight different domains (abstracts, books, news, poetry, recipes, Reddit, reviews, and wiki).

cate that TCD is not tailored to a single domain. Because each domain contains text produced by multiple source LLMs, these results also suggest that entropy-based reweighting transfers across generators rather than exploiting generator-specific artifacts. Complete per-domain AUROC for all baselines is provided in Appendix B (Table 7).

Hybrid Text Detection. Hybrid text, where passages mix human-written and LLM-generated content, poses a harder challenge than fully LLM-generated detection. We evaluate on HART (Bao et al., 2025), which defines three levels of increasing difficulty: Level-1 detects any AI involvement, Level-2 distinguishes human-authored from AI-authored base content, and Level-3 iden-

tifies fully AI-generated text (Table 2). TCD improves all three base detectors at every level. The largest gain occurs at Level-3 (+4.06% AUROC for Fast-DetectGPT+TCD), where fully LLM-generated passages exhibit the strongest contextual regularities. Level-2 remains the hardest setting for all methods (69.00% baseline AUROC for Fast-DetectGPT), yet TCD still yields a +2.93% improvement, indicating that emphasizing high-dependency tokens helps even when surface expression has been partially humanized. Detailed per-domain breakdowns across writing, essay, arXiv, and news are provided in Appendix B (Table 8).

Robustness on Adversarial Attacks. We evaluate on the adversarial attack subset of

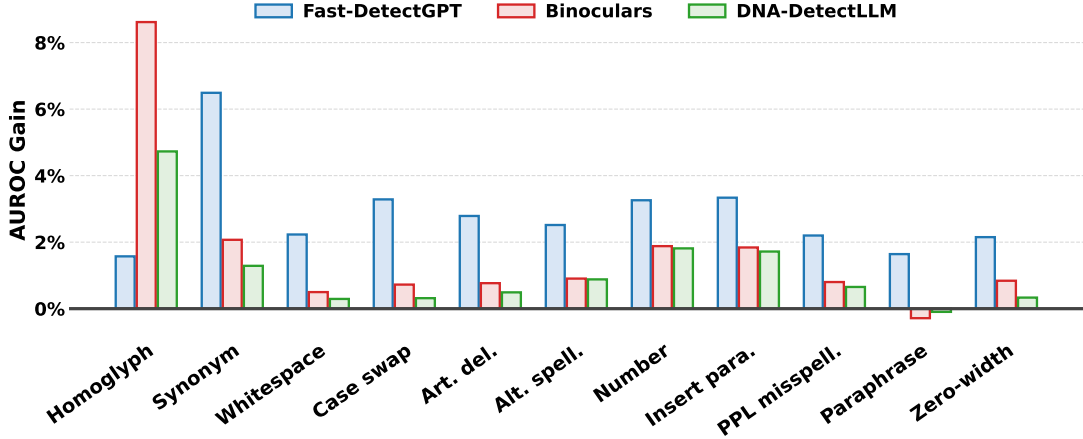


Figure 4: AUROC gain (%) of TCD over baseline detectors under 11 adversarial attacks on RAID. "AUROC Gain" is calculated by (Ours - Base).

Methods	Level-1	Level-2	Level-3
Fast-DetectGPT	81.31	69.00	84.59
+TCD (Ours)	83.02	71.93	88.65
(Diff)	1.71	2.93	4.06
Binoculars	81.68	70.10	86.43
+TCD (Ours)	83.08	72.05	89.08
(Diff)	1.40	1.95	2.65
DNA-DetectLLM	81.86	71.43	87.28
+TCD (Ours)	83.80	72.81	89.05
(Diff)	1.94	1.38	1.77

Table 2: Comparison of detection performance (Average AUROC, %) across different writing levels. Best results are **bolded**.

RAID (Dugan et al., 2024), covering 11 perturbation types. TCD improves AUROC over all three base detectors on most attack types, with the largest gains under homoglyph substitution and synonym swap (+8% and +6%, respectively) and modest but positive gains (+0.3%–+3.5%) for other surface-level perturbations. The main exception is paraphrase, where TCD slightly hurts Binoculars and yields near-zero change for DNA-DetectLLM, suggesting semantic-preserving rewrites can partially align with the reweighting signal. Overall, TCD consistently improves adversarial robustness in the majority of attacks scenarios.

Robustness on Different Text Lengths. Detection performance often degrades on short passages where fewer token-level statistics are available. We truncate RAID samples to target lengths of 40, 80, 120, 160, 200, 240, and 280 words and report the AUROC gain of TCD over each baseline (Figure 5). TCD yields positive gains across the entire length

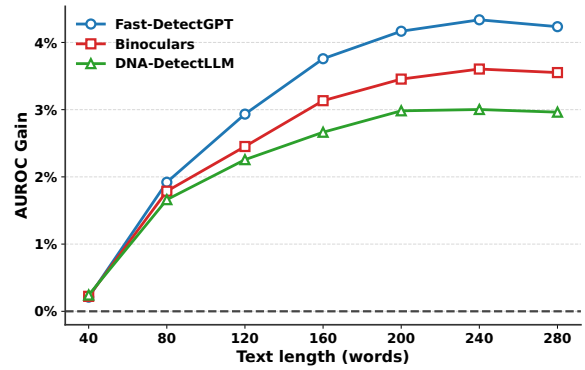


Figure 5: AUROC gain (%) of TCD over baseline detectors at varying text lengths (words) on RAID.

Model series	Fast.	Bino.	DNA.
Llama-3.2-1B family	3.13	2.32	2.23
Llama-3.2-3B family	2.28	1.57	1.79
Gemma-2-2B family	1.97	1.46	2.38

Table 3: Proxy-model sensitivity: AUROC gain (%) of TCD over the baseline detector, averaged over RealDet, RAID, and M4.

range for all three detectors. The improvement is most pronounced at moderate lengths (around 120–200 words), where enough tokens exist for meaningful entropy-based weighting yet uniform aggregation still dilutes the signal. Even at 40 words, TCD retains a gain, showing that selecting informative positions helps when the total number of tokens is limited.

Impact of Different Surrogate Models. TCD relies on attention entropy from a surrogate model; we therefore study whether its benefit persists under different model families. Besides the de-

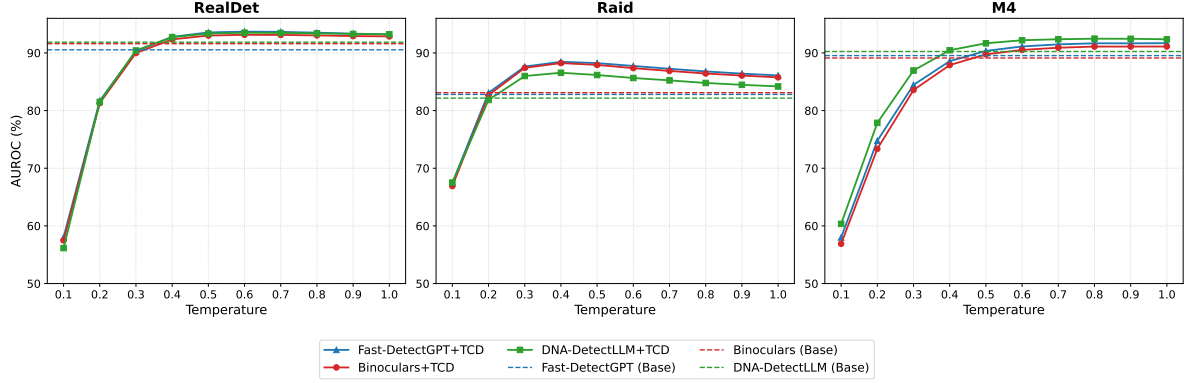


Figure 6: Sensitivity of TCD-augmented detectors to the softmax temperature τ on RealDet, RAID, and M4.

fault Llama-3.2-1B pair (Grattafiori et al., 2024), we additionally evaluate Llama-3.2-3B (Grattafiori et al., 2024) and Gemma-2-2B (Team et al., 2024) base/instruct families on RealDet, RAID, and M4 (Table 3; dataset-level results in Appendix B, Tables 10 and 11). TCD consistently improves all three detectors under every tested family, with average AUROC gains ranging from 1.46% to 3.13%. The default Llama-3.2-1B family yields the largest average gain, while Gemma-2-2B and Llama-3.2-3B families also show stable improvements. These findings indicate that TCD is not tied to a specific surrogate architecture, though stronger or better-aligned surrogates may yield larger gains.

Impact of Hyperparameters. The only additional hyperparameter introduced by TCD is the softmax temperature τ in Section 3.4, which controls how sharply token weights concentrate on low-entropy positions. Figure 6 plots AUROC on RealDet, RAID, and M4 as τ varies from 0.1 to 1.0. Performance remains largely stable across most of the search range, with TCD-augmented detectors consistently outperforming their base counterparts. Very low temperatures can over-sharpen the weight distribution by focusing on only a few tokens, whereas $\tau = 0.7$ already yields strong results and serves as our default. This mild sensitivity suggests that TCD does not require extensive hyperparameter tuning in practice.

4.4 Intervention Analysis of Weighting

Table 1 and Figure 1 show that low-entropy tokens correlate with stronger human/LLM score separation, but do not by themselves establish that these tokens *cause* the detection gain. To move from correlation to causation, we intervene on the tokens, the weights, and the attention sources used by

Configuration	Fast.	Bino.	DNA.
Base detector (no TCD)	87.63	87.94	88.09
Base detector +TCD	90.76	90.26	90.32
(a) Mask 30% low-entropy tokens	79.00	83.40	83.99
(b) Mask 30% high-entropy tokens	90.10	89.68	89.72
(c) Permute entropy weights	86.81	87.10	87.52
(d) Invert entropy weights	81.40	84.05	84.74
(e) Ablate 30% low-entropy heads	90.59	90.09	90.20
(f) Ablate low-entropy layers	88.35	88.47	88.90

Table 4: Intervention analysis of attention-entropy weighting. AUROC is averaged over RealDet, RAID, and M4. Settings (a)–(b) drop tokens and aggregate the rest uniformly; (c)–(d) keep all tokens but perturb the TCD weights; (e)–(f) remove low-entropy heads or DLS-selected layers before computing weights.

TCD (Table 4). Settings (a)–(f) are also alternative non-uniform aggregation schemes under otherwise identical conditions, so they isolate whether the gain is specific to attention-entropy weighting.

Token masking. Dropping 30% of tokens and uniformly aggregating the rest has opposite effects. Removing the lowest-entropy tokens (a) collapses AUROC well below the uniform base (Fast-DetectGPT: 87.63 $\rightarrow$ 79.00), whereas removing the highest-entropy tokens (b)—i.e., keeping only low-entropy ones—already approaches full TCD (90.10) without any reweighting.

Weight perturbation. Randomly permuting the entropy-based weights (c) returns performance to the base level (86.81), and inverting them (d) drops well below it (81.40). The gain therefore comes from up-weighting low-entropy tokens, not from arbitrary non-uniform weighting.

Head and layer ablation. Removing the lowest-entropy attention heads (e) barely changes AUROC, indicating that the signal is redundant across heads.

<i>Variant</i>		Fast.	Bino.	DNA.
<i>Cond.1</i>	<i>Cond.2</i>			
Uniform (no TCD)		87.63	87.94	88.09
✗	✗	89.48	89.30	89.53
✗	✓	90.29	89.86	89.97
✓	✗	90.40	89.99	90.03
✓	✓	90.76	90.26	90.32

Table 5: Ablation of token reweighting and of Condition 1 and Condition 2 in Dynamic Layer Selection. The uniform row aggregates token statistics without TCD. All remaining rows apply reweighting and differ only in the selected layers. AUROC is averaged over RealDet, RAID, and M4.

Ablating the DLS-selected low-entropy layers (f) removes most of the TCD gain ($90.76 \rightarrow 88.35$), localizing the causal source to the layers that DLS selects.

None of (a)–(f) recovers full TCD. Low-entropy tokens and the layers that produce them therefore causally drive detection, and the improvement is attributable specifically to attention-entropy weighting rather than to non-uniform aggregation in general.

4.5 Ablation Studies

We ablate DLS Conditions 1 and 2 individually and jointly on RealDet, RAID, and M4 (Table 5); all TCD variants apply the attention entropy-based token reweighting of Section 3.4 and differ only in the selected layers. Without reweighting, layer selection cannot affect the score: tokens are aggregated uniformly, which is exactly the base detector in Table 1 (87.63% average AUROC for Fast-DetectGPT). Turning on reweighting over all layers (both conditions off) accounts for most of the gain ($87.63 \rightarrow 89.48$, +1.85), and adding DLS on top contributes a further complementary improvement ($89.48 \rightarrow 90.76$, +1.28). Either condition alone already beats using all layers, and applying both achieves the best performance. The same pattern holds for Binoculars and DNA-DetectLLM. Reweighting is therefore the primary driver and DLS a complementary refinement, and the two selection criteria are complementary rather than redundant.

Methods	Time (s)
Fast-DetectGPT	0.070
+TCD (Ours)	0.145
Binoculars	0.372
+TCD (Ours)	0.468
DNA-DetectLLM	0.669
+TCD (Ours)	0.720

Table 6: Time efficiency analysis on the RAID benchmark. Each entry reports the average detection latency in seconds per sample (s/sample), measured with batch size = 1 after truncating texts to 300 tokens.

4.6 Efficiency Analysis

To quantify the inference overhead of TCD, we measure average detection latency on the RAID benchmark (Dugan et al., 2024) under a balanced setting: 500 human and 500 LLM texts are randomly sampled, truncated to 300 tokens, and evaluated with batch size of 1. As shown in Table 6, TCD introduces additional computation, increasing latency by 0.05–0.10 s per sample across the three backbones, yet all variants remain below one second per sample. The extra cost is incurred in the same surrogate forward pass already used to obtain token-level statistics, with no additional model call or parameters, so the absolute overhead stays roughly constant while the relative increase shrinks for heavier detectors ($2.07\times$ for Fast-DetectGPT, $1.26\times$ for Binoculars, and $1.08\times$ for DNA-DetectLLM). This latency is sufficient for real-time or interactive detection scenarios such as online content screening and writing-assistant feedback, where users expect immediate responses without offline retraining.

5 Conclusion

We show that zero-shot detection signal is concentrated at token positions of high contextual dependency, quantified by attention entropy. TCD turns this observation into a plug-and-play reweighting module that integrates into existing detectors without retraining, and yields consistent gains across benchmarks, domains, adversarial attacks, and hybrid-text settings. Our results suggest that not all token positions contribute equally, and that emphasizing high-dependency positions offers a simple way to strengthen zero-shot detectors without changing their core scoring functions.

Limitations

TCD computes attention entropy from the surrogate model that the detector runs locally, and therefore requires this surrogate to be a Transformer whose per-layer attention weights are accessible. This holds for the open-source surrogates used by existing zero-shot detectors, where attention weights and next-token logits come from the same forward pass, but not when detection relies solely on the logprobs returned by a commercial API.

Additionally, computing token-level attention entropy during the forward pass also adds inference cost. The overhead is small in practice, since it requires no additional model call or parameters (Section 4.6), but it may still matter under strict latency budgets.

Acknowledgments

We sincerely thank the reviewers for their valuable comments, which have greatly improved the quality of this paper. This work is supported by the National Natural Science Foundation of China (62376033 and 62232006) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Nihel Fatima Baarir and Abdelhamid Djeflal. 2021. Fake news detection using machine learning. In *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH)*, pages 125–130. IEEE.
- Guangsheng Bao, Lihua Rong, Yanbin Zhao, Qiji Zhou, and Yue Zhang. 2025. Decoupling content and expression: Two-dimensional detection of ai-generated text. *arXiv preprint arXiv:2503.00258*.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pages 24814–24836.
- Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.
- Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margaret Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.
- Nicholas Diakopoulos, Hannes Cools, Charlotte Li, Natali Helberger, Ernest Kung, Aimee Rinehart, and Lisa Gibbs. 2024. Generative ai in journalism: The evolution of newswork and ethics in a generative information ecosystem. In *AI Communication*. IOS Press.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492.
- Damian Okaibedi Eke. 2023. Chatgpt and the rise of generative ai: Threat to academic integrity? *Journal of Responsible Technology*, 13:100060.
- Simon Friederich and Jonathan Symons. 2023. Norms for academic writing in the era of advanced artificial intelligence. *Digital Society*, 2(3):48.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pages 111–116.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. 2024. Biscoper: Ai-generated text detection by checking memorization of preceding tokens. *Advances in Neural Information Processing Systems*, 37:104065–104090.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Forty-first International Conference on Machine Learning*.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.

- Hazem Ibrahim, Fengyuan Liu, Rohail Asim, Balaraju Battu, Sidahmed Benabderrahmane, Bashar Alhafni, Wifag Adnan, Tuka Alhanai, Bedoor AlShebli, Riyadh Baghdadi, et al. 2023. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific reports*, 13(1):12187.
- Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. 2025. M-rangedetector: Enhancing generalization in machine-generated text detection through multi-range attention masks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8971–8983.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in neural information processing systems*, 36:27469–27500.
- Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649.
- Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.
- Shixuan Ma and Quan Wang. 2024. Zero-shot detection of llm-generated text using token cohesiveness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17538–17553.
- Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2026. Not all tokens are created equal: Perplexity attention weighted networks for ai-generated text detection. *Information Fusion*, 125:103465.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1389–1403.
- Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2025. [Survey on plagiarism detection in large language models: The impact of ChatGPT and gemini on academic integrity](#). *Journal of Academic Ethics*, 23:1137–1170.
- Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412.
- Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407.
- Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pages 62143–62156.
- Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. 2025. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12298–12319.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. 2026a. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. *Advances in Neural Information Processing Systems*, 38:1745–1768.

Xiaowei Zhu, Yubing Ren, Fang Fang, Shi Wang, Yanan Cao, and Li Guo. 2026b. Exons-detect: Identifying and amplifying exonic tokens via hidden-state discrepancy for robust ai-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26324–26336.

A Dataset Details

This section describes the benchmarks used in our experiments. We follow the balanced sampling practice used in recent zero-shot detection studies (Ma and Wang, 2024; Bao et al., 2024; Zhu et al., 2026a): for each dataset, we randomly draw equal numbers of human-written and LLM-generated passages unless noted otherwise.

A.1 RealDet

RealDet (Zhu et al., 2025) is a large-scale bilingual benchmark built to reflect realistic LLM-generated text detection scenarios. The full corpus contains over 836k raw passages covering 15 textual domains and outputs from 22 LLMs, together with more than 106k human-written texts in both English and Chinese. LLM-generated samples are collected under several generic prompt types (continuation, topic-driven writing, and question answering), and the release also includes adversarial variants produced by paraphrasing and editing attacks. DNA-DetectLLM (Zhu et al., 2026a) evaluates on RealDet by sampling 2,000 balanced examples from the public release; we adopt the same protocol and use 1,000 human-written and 1,000 LLM-generated passages for the main results in Table 1.

A.2 RAID

RAID (Dugan et al., 2024) is a shared robustness benchmark designed to expose weaknesses that simpler corpora often miss. It contains more than 6 million generations from 11 LLMs across eight domains: arXiv abstracts, book summaries, news, poetry, recipes, Reddit posts, reviews, and Wikipedia. For each human-written seed document, RAID creates paired outputs under four decoding strategies (greedy and sampling, with and without repetition penalty) and eleven adversarial perturbations (Dugan et al., 2024), including

paraphrase (Krishna et al., 2023), synonym swap, homoglyph substitution, misspelling, whitespace insertion, and zero-width characters. We use a balanced subset of 1,000 human-written and 1,000 LLM-generated texts for the primary evaluation in Table 1. Additional analyses draw on RAID’s multi-domain partition (Table 7), its adversarial attack subset (Figure 4), and length-truncated variants (Figure 5).

A.3 M4

M4 (Wang et al., 2024) is a multi-generator, multi-domain, and multi-lingual corpus for black-box LLM-generated text detection, where the source LLM is unknown at test time. Human-written and LLM-generated texts are paired across six generators (GPT-4 (Achiam et al., 2023), ChatGPT, GPT-3.5, Cohere, Dolly-v2, and BLOOMz) (Wang et al., 2024), seven domains (Wikipedia, WikiHow, Reddit, question answering, news, academic abstracts, and peer reviews), and seven languages (Arabic, Bulgarian, Chinese, English, Indonesian, Russian, and Urdu). Prior work shows that detectors often struggle to generalize to unseen domains or generators on M4, making it a strong test of scoring-based methods that do not rely on training labels. Following the same balanced sampling strategy as on RealDet and RAID, we evaluate on the multilingual test partition with 1,000 human-written and 1,000 LLM-generated passages.

A.4 HART

HART (Bao et al., 2025) (Hierarchical AI Risk in Text Creation) extends binary detection to settings where human and AI contributions are mixed. Each sample is labeled by AI participation type: purely human (Type 0), human content with AI-refined expression (Type 1), AI content rewritten to appear human-like (Type 2), or fully AI-generated text (Type 3). The benchmark spans four domains (student essays, arXiv introductions, creative writing, and news) and multiple languages, with texts produced by six recent LLMs including GPT-3.5-Turbo, GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, LLaMA 3.3-70B-Instruct, and Qwen 2.5-72B-Instruct (Bao et al., 2025). HART defines three nested binary tasks: Level-1 detects any AI involvement (Type 0 vs. Types 1–3), Level-2 distinguishes human-authored from AI-authored content (Types 0–1 vs. Types 2–3), and Level-3 identifies fully AI-generated text (Types 0–2 vs. Type 3). We report aggregated results across levels in Table 2

Methods	abstracts	books	news	poetry	recipes	reddit	reviews	wiki
Likelihood	76.09	79.72	70.98	76.72	67.29	78.45	75.45	71.48
LogRank	77.42	81.52	72.30	78.21	67.44	78.85	76.77	72.45
Entropy	69.61	79.62	70.58	78.82	66.47	77.00	74.05	70.06
LRR	80.32	85.80	76.98	81.72	65.55	80.52	80.07	75.77
Fast-DetectGPT	82.68	83.56	74.42	82.53	71.61	83.68	82.46	79.70
+TCD (Ours)	83.21	86.84	77.73	87.99	74.66	85.68	83.95	80.06
(diff)	0.53	3.28	3.31	5.46	3.05	2.00	1.49	0.36
Binoculars	83.24	84.34	75.81	83.11	73.41	83.45	82.65	79.09
+TCD (Ours)	83.85	86.64	77.14	87.43	75.60	85.64	83.26	80.60
(diff)	0.61	2.30	1.33	4.32	2.19	2.19	0.61	1.51
DNA-DetectLLM	83.25	85.10	76.90	84.49	72.23	83.49	83.41	79.39
+TCD (Ours)	83.51	86.83	77.63	87.57	72.75	85.39	83.64	80.97
(diff)	0.26	1.73	0.73	3.08	0.52	1.90	0.23	1.58

Table 7: Per-domain detection performance (AUROC) on the multi-domain partition of RAID. “+TCD” denotes applying our framework to the base detector. Best results per domain among TCD-augmented methods are **bolded**.

and provide the full domain-wise breakdown in Appendix B (Table 8).

B Additional Experiments Results

This appendix provides complete numerical results that complement the analyses in Section 4.3. Due to space constraints, the main paper reports summary figures and aggregated scores; here we include per-domain and per-dataset breakdowns for multi-domain evaluation, TCD applied to simpler zero-shot detectors, hybrid text detection, surrogate-model sensitivity, DLS robustness across architectures and domains, and impact of hyperparameters.

B.1 Multi Domains and Multi Source Models

Table 7 reports AUROC (%) on the multi-domain partition of RAID across eight genres. Each domain contains LLM-generated text from multiple source LLMs, making this setting a joint test of domain transfer and generator diversity. TCD improves Fast-DetectGPT, Binoculars, and DNA-DetectLLM in all eight domains without exception. The largest Fast-DetectGPT gain occurs on poetry (+5.46%), where stylistically constrained text may produce more regular contextual patterns that TCD can exploit; books (+3.28%) and news (+3.31%) also show substantial improvements. Gains are smaller but still positive on wiki (+0.36%) and reviews (+1.49%), domains where baseline detectors are already relatively strong. Across probability-

based baselines (Likelihood, LogRank, Entropy, LRR), performance varies considerably by domain. For example, Entropy drops to 66.47% on recipes, whereas TCD-augmented multi-model detectors maintain more balanced performance across genres. These results confirm that contextual-dependency weighting is not overfit to a single text type and remains effective when both domain and source model vary.

B.2 Generalization to Simpler Zero-Shot Detectors

The main experiments apply TCD to Fast-DetectGPT, Binoculars, and DNA-DetectLLM because gains on the strongest existing methods are the most informative in practice. TCD, however, only assumes that the base detector aggregates token-level statistics, so it should also apply to simpler scoring functions. We therefore attach TCD to Likelihood, LogRank, and Entropy from GLTR (Gehrmann et al., 2019) and to LRR (Su et al., 2023), using the same protocol as Table 1 and reporting average AUROC over RealDet, RAID, and M4 (Table 9).

TCD improves all four detectors with no degradation (Likelihood +0.58, LogRank +0.26, Entropy +0.41, LRR +0.27). The uniformly positive results confirm that TCD generalizes beyond the three state-of-the-art detectors in the main text to simpler zero-shot methods that also aggregate token-level statistics.

Detector	Level-3				Level-2				Level-1			
	Writing	Essay	ArXiv	News	Writing	Essay	ArXiv	News	Writing	Essay	ArXiv	News
Fast-DetectGPT	85.59	84.60	84.61	83.57	70.64	70.01	70.40	64.94	77.75	86.92	80.81	79.76
+TCD (Ours)	89.94	86.83	89.11	88.73	75.26	69.95	72.99	69.50	80.71	88.59	81.74	81.05
(diff)	4.35	2.23	4.50	5.16	4.62	-0.06	2.59	4.56	2.96	1.67	0.93	1.29
Binoculars	86.88	87.28	86.07	85.50	71.46	71.27	70.79	66.86	77.93	87.68	80.85	80.26
+TCD (Ours)	90.14	87.51	89.54	89.13	75.04	70.15	73.05	69.96	80.73	88.92	81.83	80.85
(diff)	3.26	0.23	3.47	3.63	3.58	-1.12	2.26	3.10	2.80	1.24	0.98	0.59
DNA-DetectLLM	88.32	88.55	86.16	86.09	74.02	72.74	70.82	68.13	79.10	87.39	81.10	79.85
+TCD (Ours)	90.85	87.68	88.74	88.91	76.82	70.95	72.84	70.63	82.22	89.13	82.44	81.42
(diff)	2.53	-0.87	2.58	2.82	2.80	-1.79	2.02	2.50	3.12	1.74	1.34	1.57

Table 8: Per-domain detection performance (AUROC) on HART across three detection levels and four domains. Best results are **bolded**.

Detector	Base	+TCD (Ours)
Likelihood	84.81	85.39
LogRank	85.76	86.02
Entropy	83.09	83.50
LRR	85.87	86.14

Table 9: Detection performance (AUROC, %) of simpler zero-shot detectors with and without TCD, averaged over RealDet, RAID, and M4. Best results are **bolded**.

B.3 Hybrid Text Detection

Table 8 breaks down HART results by detection level and domain (writing, essay, arXiv, and news). Level-3 (fully AI-generated) is the easiest setting overall: Fast-DetectGPT+TCD reaches 89.94% on writing and 89.11% on arXiv, with gains of +4.35 and +4.50 over the respective baselines. Level-2 (distinguishing human-authored from AI-authored base content) remains the most challenging, especially on news, where Fast-DetectGPT improves from 64.94% to 69.50% (+4.56) with TCD, which is the largest single domain-level gain in this table. Level-1 (detecting any AI involvement) shows consistent but moderate improvements; for instance, DNA-DetectLLM+TCD reaches 89.13% on essay (+1.74 over baseline). Notably, TCD helps on arXiv and news at all three levels, suggesting that the benefit of emphasizing high-dependency tokens generalizes beyond creative writing domains to more formal scientific and journalistic text. Across all 12 domain and level combinations, TCD never reduces performance relative to the corresponding base detector.

B.4 Impact of Different Surrogate Models

Alternative surrogate families. Tables 10 and 11 report dataset-level AUROC when replacing the

default Llama-3.2-1B surrogate pair with the Llama-3.2-3B and Gemma-2-2B instruct/base families; all other settings remain identical to the main evaluation. Under both alternatives, TCD consistently improves all three base detectors on RealDet, RAID, and M4. With Llama-3.2-3B-Instruct/Llama-3.2-3B, Fast-DetectGPT+TCD raises average AUROC from 87.28% to 89.56% (+2.28), with the largest gains on RealDet (+2.50) and RAID (+3.04); Binoculars+TCD and DNA-DetectLLM+TCD gain +1.57 and +1.79 on average, respectively. With Gemma-2-2B-it/Gemma-2-2B, Fast-DetectGPT+TCD improves from 88.34% to 90.31% (+1.97), with a notable +2.33 gain on M4; DNA-DetectLLM+TCD benefits even more on M4 (+5.84), indicating that entropy-based reweighting can compensate when a base detector underperforms under a given surrogate. Across both families, TCD never degrades any detector and transfers without relying on architecture-specific attention patterns of a single checkpoint.

Official settings of each detector. Our main experiments share one surrogate family across detectors, whereas Fast-DetectGPT, Binoculars, and DNA-DetectLLM each adopt a larger, detector-specific surrogate in their official settings. To confirm that the reported gains are not an artifact of our choice, we re-evaluate each detector under its official setting, keeping everything else identical to Table 1. As shown in Table 12, TCD still improves every detector, so the gains also hold under each detector’s official configuration.

B.5 Robustness of Dynamic Layer Selection

Section 3.3 argues that the two DLS conditions follow from properties shared by Transformer LLMs and are therefore not tied to a specific surrogate or

Methods	RealDet	RAID	M4
Fast-DetectGPT	89.71	83.62	88.52
+TCD (Ours)	92.21	86.66	89.80
Binoculars	90.60	83.87	88.14
+TCD (Ours)	91.79	86.37	89.17
DNA-DetectLLM	90.75	82.25	88.29
+TCD (Ours)	91.91	84.52	90.21

Table 10: Detection performance (AUROC) using Llama-3.2-3B-Instruct and Llama-3.2-3B as the surrogate pair. Best results are **bolded**.

Methods	RealDet	RAID	M4
Fast-DetectGPT	90.99	85.24	88.78
+TCD (Ours)	93.20	86.61	91.11
Binoculars	88.17	83.74	87.16
+TCD (Ours)	89.92	84.03	89.51
DNA-DetectLLM	87.68	83.22	84.10
+TCD (Ours)	88.83	83.38	89.94

Table 11: Detection performance (AUROC) using Gemma-2-2B-it and Gemma-2-2B as the surrogate pair. Best results are **bolded**.

text distribution. We verify this by repeating the ablation of Table 5 across surrogate architectures (Table 13) and across the eight domains of RAID (Table 14).

Across surrogate architectures. With Llama-3.2-1B, Llama-3.2-3B, and Gemma-2-2B as the surrogate, the full DLS ($\checkmark/\checkmark$) is the best configuration in every column and always improves over the base detector, and the gain grows monotonically as conditions are added: either condition alone already improves over aggregating attention entropy across all layers ($\times/\times$), and combining both is strictly best. The two conditions therefore remain complementary under all three architectures, and neither relies on the attention pattern of a single checkpoint.

Across domains. Within each of the eight RAID domains, evaluated with Llama-3.2-1B as the surrogate, the same ordering holds: the full DLS yields the largest gain in all eight domains for Fast-DetectGPT and Binoculars and in seven of eight for DNA-DetectLLM. The criteria are thus not tuned to a particular text distribution.

Methods	AUROC
GPT-Neo-2.7B / GPT-J-6B	
Fast-DetectGPT	85.99
+TCD (Ours)	87.21
Falcon-7B-Instruct / Falcon-7B	
Binoculars	88.72
+TCD (Ours)	90.06
DNA-DetectLLM	90.16
+TCD (Ours)	91.22

Table 12: Detection performance (AUROC, %) averaged over RealDet, RAID, and M4 when each detector uses its official surrogate setting. Best results are **bolded**.

Surrogate	Cond.1	Cond.2	Fast.	Bino.	DNA.
Llama-3.2-1B	$\times$	$\times$	1.85	1.36	1.44
	$\times$	$\checkmark$	2.66	1.92	1.88
	$\checkmark$	$\times$	2.77	2.05	1.94
	$\checkmark$	$\checkmark$	3.13	2.32	2.23
Llama-3.2-3B	$\times$	$\times$	1.68	1.15	1.56
	$\times$	$\checkmark$	1.89	1.28	1.68
	$\checkmark$	$\times$	2.02	1.39	1.62
	$\checkmark$	$\checkmark$	2.28	1.57	1.79
Gemma-2-2B	$\times$	$\times$	1.42	1.01	1.34
	$\times$	$\checkmark$	1.60	1.26	1.38
	$\checkmark$	$\times$	1.73	1.34	2.09
	$\checkmark$	$\checkmark$	1.97	1.46	2.38

Table 13: AUROC gain (%) of TCD over the base detector under different DLS variants and surrogate families, averaged over RealDet, RAID, and M4.

B.6 Impact of Hyperparameters

TCD introduces a single additional hyperparameter, the softmax temperature τ in Section 3.4, which controls how sharply token weights concentrate on low-entropy positions. Figure 6 in the main paper summarizes the trend on RealDet, RAID, and M4; Tables 15, 16, and 17 report the full AUROC values for $\tau \in \{0.1, 0.2, \dots, 1.0\}$. Across all three benchmarks and base detectors, very small τ (e.g., 0.1) degrades performance because the weighting collapses onto too few tokens, whereas AUROC rises quickly once $\tau \geq 0.3$ and remains largely stable for $\tau \in [0.5, 1.0]$. We therefore use $\tau = 0.7$ as the default in all main experiments.

C More Related Work

Prior work has also questioned uniform token aggregation, but differs from TCD in the principle used to decide a token’s importance. PAWN (Miralles-González et al., 2026) is supervised: it trains lightweight MLPs over the surrogate’s last hidden states and token positions to learn

Detector	Cond.1	Cond.2	abstracts	books	news	poetry	recipes	reddit	reviews	wiki
Fast-DetectGPT	✗	✗	0.09	1.01	1.60	3.41	2.30	1.64	0.51	0.13
	✗	✓	0.13	2.50	2.49	4.66	3.00	1.84	1.02	0.30
	✓	✗	0.20	2.54	2.60	4.64	2.12	1.77	1.24	0.20
	✓	✓	0.53	3.28	3.31	5.46	3.05	2.00	1.49	0.36
Binoculars	✗	✗	0.08	0.36	0.21	2.60	1.43	1.87	0.02	0.06
	✗	✓	0.11	1.56	0.46	3.71	2.14	2.13	0.05	0.70
	✓	✗	0.36	1.63	0.66	3.59	1.26	1.95	0.36	0.79
	✓	✓	0.61	2.30	1.33	4.32	2.19	2.19	0.61	1.51
DNA-DetectLLM	✗	✗	0.04	0.15	0.04	1.76	0.11	1.63	0.04	0.69
	✗	✓	0.09	1.23	0.07	2.59	0.47	1.93	0.09	1.05
	✓	✗	0.06	1.22	0.26	2.35	0.04	1.73	0.08	1.13
	✓	✓	0.26	1.73	0.73	3.08	0.52	1.90	0.23	1.58

Table 14: Per-domain AUROC gain (%) of TCD over the base detector under different DLS variants on the multi-domain partition of RAID, with Llama-3.2-1B as surrogate.

Methods	Temperature									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Fast.+TCD	58.02	81.71	90.29	92.79	93.55	93.69	93.65	93.51	93.34	93.21
Bino.+TCD	57.48	81.30	89.98	92.35	93.01	93.14	93.12	93.04	92.92	92.87
DNA.+TCD	56.15	81.49	90.42	92.73	93.36	93.50	93.46	93.39	93.29	93.23

Table 15: AUROC (%) on **RealDet** with different temperature hyperparameters.

Methods	Temperature									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Fast.+TCD	67.44	83.09	87.63	88.46	88.23	87.74	87.25	86.79	86.41	86.09
Bino.+TCD	66.91	82.69	87.41	88.24	87.92	87.36	86.88	86.43	86.07	85.76
DNA.+TCD	67.51	81.85	85.99	86.56	86.17	85.65	85.24	84.79	84.47	84.19

Table 16: AUROC (%) on **Raid** with different temperature hyperparameters.

Methods	Temperature									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Fast.+TCD	57.99	74.76	84.46	88.57	90.35	91.12	91.48	91.67	91.66	91.69
Bino.+TCD	56.89	73.36	83.56	87.89	89.75	90.52	90.90	91.10	91.10	91.11
DNA.+TCD	60.35	77.87	86.95	90.47	91.67	92.20	92.37	92.47	92.45	92.37

Table 17: AUROC (%) on **M4** with different temperature hyperparameters.

per-token weights, which reweight next-token-distribution metrics before aggregation. Exons-Detect (Zhu et al., 2026b) is training-free but relies on two proxy models, treating tokens whose hidden-state discrepancy between the two exceeds a threshold as carrying stronger source signal. TCD instead derives token importance from a single surrogate’s token contextual dependency: low entropy indicates that the prediction relies on a concentrated context, a formulaic and pretraining-familiar pattern where LLM-generated and human-written text diverge most. Attention has also been used for detection by TDA (Kushnareva et al., 2021), but at a different granularity and for a different purpose: it represents the whole attention map as a graph

and feeds its topological features to a supervised classifier, whereas TCD reads a per-token statistical property of attention as a training-free importance signal for reweighting.

D Impact Statement

This study acknowledges potential societal impacts of TCD. We caution against its misuse for mass surveillance, unjustified censorship, or discriminatory application, and advocate for responsible deployment that respects individual privacy and freedom of expression.