TIDE: ZERO-SHOT LLM-GENERATED TEXT DETECTION VIA DIRECTIONAL GENERATION ASYMMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

With the rapid advance and wide use of large language models (LLMs), the text they generate is flooding into classrooms, publications, online forums and beyond. Telling it apart from human writing has become a practical necessity as concerns over misuse and attribution grow. Many existing detectors measure how predictable a text is to a surrogate LLM, reading it from left to right (*i.e.*, in forward order), the order in which it was produced. However, we observe that adding a reading in the opposite direction (*i.e.*, in reverse order) reveals a complementary and even stronger signal for detection. Specifically, reversal makes every text harder for the surrogate to predict, but LLM-generated text loses disproportionately more predictability than human text. We call this effect *directional generation asymmetry* and build our detector, TIDE, on it. TIDE computes the cross-entropy losses of a text and its reversal in one batched call to the frozen surrogate and combines the two losses into a single directional asymmetry score. On our new dataset of $225{,}000$ documents spanning five domains, five frontier LLMs, six languages and six adversarial attacks, TIDE outperforms nine state-of-the-art zero-shot baselines, achieving an average AUROC of 0.978, 4.7% higher than the best baseline. At a practical operating point of a 0.1% false-positive rate, TIDE nearly doubles the true-positive rate of the best baseline, from 41% to 79%.

1 INTRODUCTION

Large language models (LLMs) have been widely adopted as writing tools in both personal and professional contexts (Liang et al., 2025; Chatterji et al., 2025). Studies show that they improve the quality and speed of both creative writing (Doshi & Hauser, 2024; Dell'Acqua et al., 2023) and professional writing (Noy & Zhang, 2023). However, when the provenance of a text matters, the undisclosed use of LLMs raises serious concerns. For example, generated coursework threatens academic integrity (Cotton et al., 2024), 22.5% of ICML 2026 reviewers barred from using LLMs admitted using them (Kim et al., 2026), and machine-generated articles have spread rapidly on misinformation news sites (Hanley & Durumeric, 2024). To prevent such misuse, third parties such as educators, publishers and online forums need to be able to tell LLM-generated text from human writing. Regulators have also moved to make AI-generated content identifiable. The EU AI Act obliges providers to mark AI-generated content in a machine-readable format (European Parliament and Council of the European Union, 2024), and China's Measures for Labeling of AI-Generated Synthetic Content, in force since September 2025, require service providers to label AI-generated text, images, audio and video (Translate, 2025).

Many existing detectors share the same paradigm. They pass a text through a frozen LLM, the *surrogate*, and compute a statistic of how well the text matches the surrogate's next-token predictions, such as how the text's likelihood changes when it is perturbed or resampled (Mitchell et al., 2023; Bao et al., 2024). Whatever the statistic, these detectors measure how predictable the text is when the surrogate reads it from left to right (*i.e.*, in forward order), the same order in which the text was produced. However, LLMs are trained to imitate human writing in exactly this order (Ouyang et al., 2022), so for text from frontier LLMs, the forward reading often leaves the two classes overlapping, as Figure 1(a) shows. Predictability in the forward order alone is therefore no longer enough to tell LLM-generated text from human writing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we observe that every text becomes harder to predict when the surrogate reads it from right to left (*i.e.*, in reverse order), but LLM-generated text loses disproportionately more predictability than human writing. We call this effect *directional generation asymmetry*. The drop in predictability itself is expected, since causal LLMs are known to struggle with reversed input (Yu et al., 2025). Its magnitude reflects how each class of text is produced. An LLM generates text autoregressively and usually picks high-probability continuations of the tokens before it, so its output aligns with the forward predictions of a causal surrogate, even one that differs from the generator, as long as both share the common regularities of language. Reversal removes this alignment, since the surrogate must then predict each token from the tokens that originally followed it. A human writer instead plans and revises a text as a whole, so human writing gains less predictability from the forward order and loses less when reversed (Jiao et al., 2025). The reverse reading therefore complements the forward one. As we show in §3, the *reverse loss*, the surrogate's cross-entropy on the reversed text, separates the two classes no better than the *forward loss* on the original text. Yet a combination of the two losses, which measures each text's drop in predictability from the forward to the reverse order, separates the classes far better than either loss alone. To the best of our knowledge, no existing detector reads a text against the direction in which it was generated.

Building on this observation, we propose TIDE, a zero-shot detector that pairs the token sequence of a text with its reversal and scores both with a frozen surrogate in one batched call. TIDE combines the resulting forward and reverse losses into a directional asymmetry score that multiplies their difference by their ratio. A higher score indicates LLM-generated text, while a lower score indicates human writing. As a result, TIDE separates the two classes of text better than state-of-the-art zero-shot detectors, especially at the low false-positive rates used in practice. Because the two sequences share one LLM inference, TIDE runs in less than twice the time of a single forward pass of the surrogate, a small fraction of the cost of detectors that regenerate the text (Yang et al., 2024; Mao et al., 2024; Sun et al., 2026). In summary, we make the following contributions:

- We identify *directional generation asymmetry*, an effect in which LLM-generated text loses disproportionately more predictability than human writing when a causal LLM reads each in reverse order, and show that it is complementary to the forward-order signals used by existing detectors.
- We propose TIDE, which reads a text in both orders in one batched call to a frozen surrogate and combines the two losses into a directional asymmetry score. Evaluated with 16 surrogates, from GPT-2 to current models, TIDE consistently outperforms the forward-order-only baseline.
- We build a dataset of $225{,}000$ documents covering five domains, five frontier LLMs, six languages and six adversarial attacks, and compare TIDE with nine state-of-the-art zero-shot baselines. TIDE achieves the best AUROC on every generator and in every language, improving the average AUROC from the best baseline's 0.934 to 0.978. It also improves the true-positive rate from the best baseline's 41% to 79% at the practical operating point of a 0.1% false-positive rate.

2 Background and Related Work

2.1 Detecting LLM-Generated Text

Methods for detecting LLM-generated text fall into two groups according to the access they need to the generator, the model that produces the text (Tang et al., 2024; Wu et al., 2025a). White/gray-box methods need access to the generator itself, typically to control its decoding. The most widely studied of them is watermarking (Liu et al., 2024b; Lalai et al., 2025; Gloaguen et al., 2025; Liang et al., 2026; Ye et al., 2026), which alters the generator's sampling so that its output carries a statistical signal that the holder of a secret key can later detect (Kirchenbauer et al., 2023; Kuditipudi et al., 2024; Dathathri et al., 2024; Hu et al., 2024; Liu et al., 2024a; Piet et al., 2025; Zhao et al., 2025; Wu et al., 2026b;c; Huo et al., 2026). In practice, embedding and verifying a watermark therefore rely on the model's provider or need an intrusion into the text generation procedure (Liu et al., 2026). A third party, such as a school or a publisher, usually sees only the text and lacks the key needed to detect a watermark. Black-box methods, which need no access to the generator, are the practical choice for such users.

Black-box methods in turn fall into three categories: statistical, supervised and zero-shot. Statistical methods compute hand-crafted features of the text, such as lexical diversity, repetition and coherence, that target known flaws of GPT-2, Grover and other early generators (Fröhling & Zubiaga,

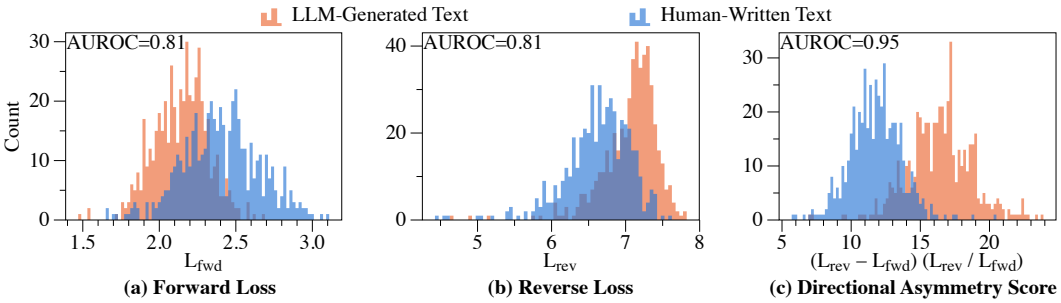


Figure 1: **Neither reading direction alone separates the classes well, but combining the two does.** (a) Forward loss, the surrogate's mean cross-entropy on the text in forward order. (b) Reverse loss, the surrogate's mean cross-entropy on the same tokens in reverse order. (c) Directional asymmetry score, which multiplies the difference of the two losses by their ratio.

2021; Gallé et al., 2021). Their accuracy falls as generators grow larger and more capable (Fröhling & Zubiaga, 2021; Uchendu et al., 2021), and they have not kept pace with frontier LLMs. Supervised methods learn from labeled examples of human and LLM-generated text, either by training a classifier on features that a language model extracts from the text (Guo et al., 2024a; Verma et al., 2024; Yu et al., 2024; Song et al., 2025; Guo et al., 2025b; 2026) or by fine-tuning a pretrained encoder or LLM into a classifier (Solaiman et al., 2019; Hu et al., 2023; Wang et al., 2024a; Guo et al., 2024b; He et al., 2025; Li et al., 2026). Such methods need labeled data for training and transfer poorly to generators, domains and languages that the data does not cover (Macko et al., 2023; Wang et al., 2024c; Dugan et al., 2024; Wang et al., 2024b; He et al., 2024; Quaremba et al., 2026).

Zero-shot methods threshold one easily interpreted score computed with a frozen surrogate LLM. They typically require no training or only lightweight threshold calibration, and thus apply to new generators, domains and languages without retraining. TIDE belongs to this family. Early methods score each token's probability, rank or entropy (Gehrmann et al., 2019; Solaiman et al., 2019; Ippolito et al., 2020), and LRR (Su et al., 2023) divides the log-likelihood by the log-rank. DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) measure the probability curvature, which DMAP (Kempton et al., 2026) re-examines and AdaDetectGPT (Zhou et al., 2025) sharpens with a learned witness function. Binoculars (Hans et al., 2024) and DNA-DetectLLM (Zhu et al., 2025) divide a perplexity by a cross-perplexity of two surrogates. Lastde (Xu et al., 2025) and SpecDetect (Luo et al., 2026) analyze the token log-probabilities as time series data. Glimpse (Bao et al., 2025) extends such statistics to proprietary models, and Markov-informed calibration (Wu et al., 2026a) smooths their token-level scores. DNA-GPT (Yang et al., 2024), Raidar (Mao et al., 2024) and D&R (Sun et al., 2026) compare the text with an LLM's continuation, rewrite or restoration of it, and Learn-to-Distance (Zhou et al., 2026) learns the distance between the text and its rewrite. Others match the text against web text (Lu et al., 2025) or add signals from its grammar and token cohesiveness (Wu et al., 2025b; Ma & Wang, 2024). All of these methods read the text only in its forward order. We compare TIDE with nine of them (§5.1).

2.2 READING TEXT IN REVERSE

How causal LLMs handle reversed text has been studied for its own sake. A model trained on reversed text predicts reversed text almost as well as a forward model predicts forward text, with a slightly higher loss (Papadopoulos et al., 2024) and, as theory implies, largely the same perplexity on each text (Luo et al., 2025). A model trained only forward, in contrast, fails on reversed input (Yu et al., 2025). The same study trains a model on both orders and uses the gap between its forward and reverse losses, which tracks text quality, to select pretraining data. Reading reversed text with a frozen LLM also yields text embeddings that complement forward ones (Lin et al., 2026), and reverse-order language models have been trained to score a prompt given its response (Yerram et al., 2024) or to complement a forward model in generation (Yin et al., 2026). Some detectors change the order of a text. Luminol-AIDetect (La Cava & Tagarelli, 2026) models the perplexity shifts of randomly shuffled text, D&R (Sun et al., 2026) scores how well a black-box LLM restores locally shuffled words, and Sun & Lv (2025) reorder the text with a dependency parser and score the sim-

ilarity between the reordered and original text. TIDE instead reverses the whole token sequence, which is deterministic and keeps every token, and compares a frozen surrogate's losses on the text and on its reversal. To the best of our knowledge, TIDE is the first detector to make this comparison.

3 MOTIVATION: READING IN BOTH DIRECTIONS

Most existing zero-shot detectors introduced in §2.1 judge a text's provenance by how well the surrogate predicts it from left to right (*i.e.*, in forward order). An intuitive and simple measure is the surrogate's mean cross-entropy, the *forward loss* L_{fwd}, which averages the cross-entropy at each token position given all the tokens before it. A lower forward loss means the text is more aligned with the surrogate's predictions. Figure 1(a) shows the forward loss of QA-forum posts written by GPT-5-mini (Singh et al., 2025) and by humans, computed with the surrogate `llama2-7b-chat` (Touvron et al., 2023). The LLM-generated posts have the lower loss, yet the two distributions overlap over half of their range, and the forward loss separates them with an AUROC of only 0.81.

This overlap is expected. The forward reading follows the order in which an LLM generates text, and LLMs are trained to imitate human writing in exactly this order (Ouyang et al., 2022). Read in this direction, text from frontier LLMs is hard to tell apart from human writing. The two classes differ, however, in how they are written. An LLM chooses each token given only the tokens before it, whereas a human writer plans and revises a text as a whole, so human writing depends not only on what precedes each word but also on what follows it (Jiao et al., 2025). **Therefore, we hypothesize that reading a text from right to left (*i.e.*, in reverse order) exposes a complementary difference that the forward reading misses.** If so, comparing the losses of a text in the two orders should separate the classes better than either loss alone.

To test this hypothesis, we compute the *reverse loss* L_{rev}, which averages the cross-entropy at each token position given all the tokens after it, fed to the surrogate in reverse order. As Figure 1(b) shows, L_{rev} is larger than L_{fwd} for every post, and the LLM-generated posts become harder to read than the human-written ones, unlike in the forward reading. **We thus observe that reversal makes every text harder to predict, but LLM-generated text loses disproportionately more predictability than human writing, an effect we call *directional generation asymmetry*.** The reverse loss alone separates the classes no better than the forward loss, with the same AUROC of 0.81, whereas a score that combines the two losses reaches 0.95, as predicted (Figure 1(c)). Comparing the two losses of the same text cancels the properties that raise or lower both, such as topic, wording and intrinsic difficulty, and keeps the part of predictability that depends on the reading direction. We name this score the *directional asymmetry score*, detail it in §4.4 and evaluate it more thoroughly in §5.5. Based on this observation, we build TIDE.

4 DESIGN OF TIDE

4.1 OVERVIEW

TIDE is a zero-shot detector that works in a black-box setting, with no access to the generator LLM and no training. Its only resource is a small surrogate language model $\mathcal{M}$. The surrogate does not need to match the generator's architecture, size or training data, and its weights stay frozen during detection. As Figure 2 shows, TIDE consists of three stages.

Stage ❶: Directional Sequence Pairing (§4.2). TIDE tokenizes the text once and pairs the *forward-order* sequence, which is the text as written, with the *reverse-order* sequence, which lists the tokens from last to first. The pair forms a single batch.

Stage ❷: Directional Loss Extraction (§4.3). TIDE reads both sequences with the surrogate in a single batched call and extracts one *directional loss* per sequence, the mean cross-entropy of each token given the tokens before it in that sequence. The *forward loss* L_{fwd} is the feature that many existing detectors build on, and the *reverse loss* L_{rev} is the feature that TIDE introduces.

Stage ❸: Directional Asymmetry Scoring (§4.4). TIDE quantifies the directional generation asymmetry observed in §3 by measuring how much harder the reverse-order sequence is for the surrogate to read than the forward-order one, as the difference $L_{\mathrm{rev}} - L_{\mathrm{fwd}}$ multiplied by the ratio $L_{\mathrm{rev}}/L_{\mathrm{fwd}}$. A higher directional asymmetry score indicates that the text is more likely LLM-generated.

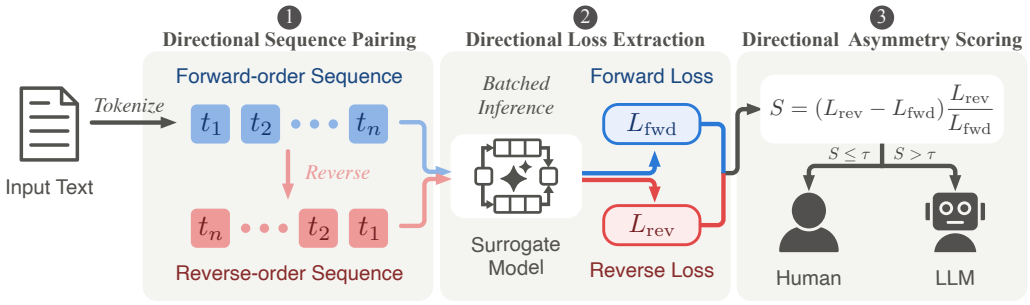


Figure 2: **Overview of TIDE. Stage ❶** pairs the tokenized input text with its reversal. **Stage ❷** computes the forward and reverse losses in one batched call to the frozen surrogate. **Stage ❸** combines the two losses into the directional asymmetry score S and labels the text with a threshold τ.

4.2 DIRECTIONAL SEQUENCE PAIRING

Given an input text x, TIDE first applies the surrogate's tokenizer, which splits the text into a sequence of tokens $\boldsymbol{t}_{\text{fwd}} = (t_1, \ldots, t_n)$. This is the forward-order sequence, in which the tokens appear in the order they were written or generated. TIDE then reverses the whole forward-order sequence and obtains the reverse-order sequence $\boldsymbol{t}_{\text{rev}} = (t_n, \ldots, t_1)$, in which every pair of neighboring tokens appears in the opposite order. The only difference between the forward-order and reverse-order sequences is the token order, and we neither change any token ID nor shuffle any block of tokens. The two sequences share the same length, and we assemble them into one inference batch for Stage ❷.

4.3 DIRECTIONAL LOSS EXTRACTION

TIDE passes the batch through the surrogate $\mathcal{M}$ in one call. The surrogate reads each sequence and, at every position, outputs a probability distribution over the next token given the tokens before that position. From these distributions, TIDE computes the mean per-token cross-entropy loss of each sequence individually,

$$L_{\text{fwd}}(\boldsymbol{t}_{\text{fwd}}) = \frac{1}{n-1} \sum_{i=2}^{n} \text{CE}(t_i,\ \boldsymbol{p}_{\mathcal{M}}(t_1, \ldots, t_{i-1}))\,, \tag{1}$$

$$L_{\text{rev}}(\boldsymbol{t}_{\text{rev}}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \text{CE}(t_i,\ \boldsymbol{p}_{\mathcal{M}}(t_n, \ldots, t_{i+1}))\,, \tag{2}$$

where $\boldsymbol{p}_{\mathcal{M}}(\cdot)$ is the probability vector that the surrogate predicts over its vocabulary $\mathcal{V}$ for the next token, and CE is the cross-entropy between the ground-truth token t_i and a predicted probability vector $\boldsymbol{p}$,

$$\text{CE}(t_i, \boldsymbol{p}) = -\sum_{v \in \mathcal{V}} \mathbb{1}[v = t_i]\, \log p_v, \tag{3}$$

where p_v is the entry of $\boldsymbol{p}$ for token v, and $\mathbb{1}[v = t_i]$ equals 1 if v is t_i and 0 otherwise. Since no token precedes it, the first token of each sequence serves only as context, and each loss averages over the other $n - 1$ tokens. L_{fwd} is the *forward loss*, in which each token is predicted from the tokens before it in the input text, and L_{rev} is the *reverse loss*, in which each token is predicted from the tokens after it in the input text. Because the surrogate was trained to read from left to right, L_{rev} is always larger than L_{fwd}, and Stage ❸ measures how much larger it is via one scalar score.

4.4 DIRECTIONAL ASYMMETRY SCORING

Stage ❸ turns the two directional losses into one scalar score that grows with the asymmetry between the reading directions. How much larger L_{rev} is than L_{fwd} can be measured in two natural ways. The difference $L_{\text{rev}} - L_{\text{fwd}}$ gives it in absolute terms, and the ratio $L_{\text{rev}}/L_{\text{fwd}}$ gives it relative to the difficulty of the forward reading. The two can rank texts differently. For example, a text with losses

of 2 and 6 and a text with losses of 4 and 8 have the same difference of 4 but ratios of 3 and 2. To combine the strengths of both, TIDE takes their product as the *directional asymmetry score*,

$$S(x) = \underbrace{(L_{\text{rev}} - L_{\text{fwd}})}_{\text{difference}} \cdot \underbrace{\frac{L_{\text{rev}}}{L_{\text{fwd}}}}_{\text{ratio}}, \tag{4}$$

so a text scores high when its reverse loss exceeds its forward loss both by a large amount and by a large factor. We ablate each factor of this score in §5.5.

Detection. A higher *directional asymmetry score* indicates that the text is more likely LLM-generated, so in deployment TIDE labels x as LLM-generated when $S(x) > \tau$ for a decision threshold τ.

5 EVALUATION

5.1 EXPERIMENTAL SETUP

Dataset. Our dataset has 225,000 documents from five domains (customer reviews and news (Wu et al., 2024), peer reviews (Yu et al., 2026), QA-forum posts (Wang et al., 2024c) and student essays (Koike et al., 2024)) and five frontier LLMs (GPT-5-mini (Singh et al., 2025), Claude Sonnet 4.5 (Anthropic, 2025), Gemini 3.1 Flash-Lite (Google DeepMind, 2026), DeepSeek-R1 (Guo et al., 2025a) and Mistral Large 3 (Mistral AI, 2025)). Each domain has 2,500 human-written English documents, each paired with one counterpart per LLM of matched topic and word count. We call each (domain, generator) pair a *cell* and evaluate on 500 test documents per class, keeping the other 2,000 as a training set for future work on detectors that require training. Test documents are also translated into Chinese, German, Arabic, Spanish and French with the Google Translate API, and the LLM-generated ones are attacked by DIPPER (Krishna et al., 2023), Chinese and German back-translation (Macko et al., 2024), DeepWordBug (Gao et al., 2018), TextFooler (Jin et al., 2020) and TextBugger (Li et al., 2019) (Appendix A).

Surrogates and baselines. Since a zero-shot detector depends on its surrogate, we evaluate every method with two surrogates from different families, `qwen3.5-4b` (Qwen Team, 2026) (the default) and `llama2-7b-chat` (Touvron et al., 2023), and evaluate TIDE with 14 more surrogates in §5.5. We compare TIDE with nine zero-shot baselines that use diverse scoring methods, including Entropy, Log-Rank (Gehrmann et al., 2019), LRR (Su et al., 2023), Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), DNA-DetectLLM (Zhu et al., 2025), Lastde (Xu et al., 2025), SpecDetect (Luo et al., 2026) and DNA-GPT (Yang et al., 2024). Each baseline follows its official implementation. All methods share the surrogate and a 1,024 input token limit, and we present the other variants of Fast-DetectGPT, Lastde and SpecDetect in Appendix B.

Metrics. We report the area under the ROC curve (AUROC) on the test set of each cell, averaged first over the generators and then over the domains. An AUROC below 0.5 means that a method is inverted and, in the score direction its authors define, ranks human-written text above LLM-generated text. We also report the true-positive rate (TPR) at fixed false-positive rates (FPRs) of 10%, 5%, 1% and 0.1% on English.

5.2 MAIN RESULTS

Table 1 reports the AUROC of every method, with English results listed per generator, other languages averaged over generators and the six attacks averaged. TIDE achieves the highest AUROC in every column on both surrogates. Its average AUROC is 0.978 on `qwen3.5-4b`, 4.7% higher than the 0.934 of the best baseline, Entropy, and 0.961 on `llama2-7b-chat`, 3.3% higher than the 0.930 of DNA-DetectLLM. We present paired significance tests of each margin in Appendix C.

Across domains. TIDE leads in every domain on `qwen3.5-4b` (Figure 3(a)). Its largest margin is on peer reviews, with 0.976 against 0.887 for Entropy, and its smallest margins are 0.008 on customer reviews and news. On QA-forum posts and student essays, it leads by 0.066 and 0.043. On `llama2-7b-chat`, TIDE leads in four of the five domains, again with its largest margin on peer reviews (0.927 against 0.825 for Fast-DetectGPT), while DNA-DetectLLM leads on QA-forum posts

Table 1: Detection AUROC of TIDE and nine zero-shot baselines on the two surrogates, macro-averaged over the five domains. The best result in each column is in bold and the second best is underlined. The last row of each block reports the improvement of TIDE over the best baseline in that column. We do not evaluate Arabic (ar) on `llama2-7b-chat`, since Llama 2 does not support this language (Touvron et al., 2023).

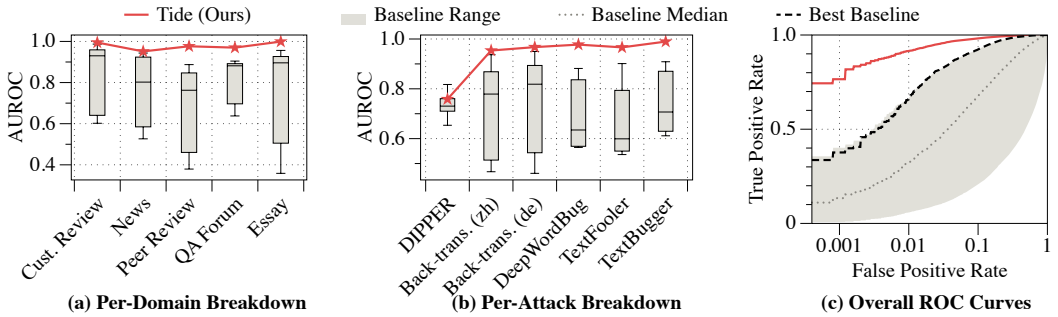
	English, per generator					Other languages						
Method	Claude	DeepSeek	Gemini	GPT	Mistral	zh	de	ar	es	fr	Attacks	Avg.
Surrogate: `qwen3.5-4b`												
Entropy	0.991	0.996	0.988	0.947	0.997	0.797	0.925	0.875	0.939	0.925	0.899	0.934
Log-Rank	0.993	0.994	0.990	0.893	0.997	0.830	0.899	0.857	0.922	0.913	0.854	0.922
LRR	0.984	0.972	0.973	0.858	0.991	0.858	0.809	0.797	0.898	0.863	0.792	0.890
Fast-DetectGPT	0.708	0.468	0.828	0.477	0.762	0.553	0.447	0.515	0.509	0.475	0.588	0.576
Binoculars	0.714	0.468	0.836	0.475	0.772	0.546	0.444	0.510	0.502	0.467	0.587	0.575
DNA-DetectLLM	0.513	0.287	0.755	0.441	0.564	0.492	0.446	0.535	0.473	0.431	0.572	0.501
Lastde	0.963	0.879	0.947	0.765	0.968	0.709	0.426	0.408	0.621	0.597	0.679	0.724
SpecDetect	0.987	0.983	0.989	0.873	0.995	0.825	0.820	0.782	0.881	0.863	0.702	0.882
DNA-GPT	0.958	0.959	0.946	0.790	0.984	0.775	0.885	0.780	0.892	0.887	0.796	0.877
TIDE (ours)	**0.998**	**1.000**	**0.999**	**0.980**	**1.000**	**0.941**	**0.986**	**0.963**	**0.983**	**0.977**	**0.936**	**0.978**
Δ vs. best baseline	+0.005	+0.004	+0.009	+0.033	+0.003	+0.083	+0.061	+0.088	+0.044	+0.052	+0.037	+0.044
Surrogate: `llama2-7b-chat`												
Entropy	0.959	0.957	0.904	0.723	0.982	0.729	0.811	—	0.858	0.826	0.716	0.846
Log-Rank	0.971	0.984	0.951	0.766	0.995	0.735	0.794	—	0.847	0.814	0.794	0.865
LRR	0.944	0.870	0.863	0.662	0.964	0.748	0.723	—	0.796	0.756	0.653	0.798
Fast-DetectGPT	0.983	0.994	0.970	0.938	0.994	0.780	0.849	—	0.901	0.909	0.939	0.926
Binoculars	0.987	0.995	0.974	0.926	0.996	0.770	0.838	—	0.896	0.902	0.934	0.922
DNA-DetectLLM	0.992	0.995	0.976	0.910	0.998	0.796	0.869	—	0.921	0.923	0.922	0.930
Lastde	0.265	0.147	0.323	0.307	0.307	0.415	0.197	—	0.378	0.298	0.401	0.304
SpecDetect	0.904	0.933	0.903	0.755	0.980	0.621	0.615	—	0.608	0.653	0.705	0.767
DNA-GPT	0.826	0.853	0.701	0.676	0.880	0.459	0.641	—	0.633	0.632	0.703	0.700
TIDE (ours)	**0.995**	**0.998**	**0.995**	**0.973**	**0.999**	**0.890**	**0.927**	—	**0.952**	**0.930**	**0.954**	**0.961**
Δ vs. best baseline	+0.003	+0.003	+0.019	+0.035	+0.001	+0.094	+0.058	—	+0.031	+0.007	+0.015	+0.031

(0.965 against 0.928). We present the per-domain and per-attack results on `llama2-7b-chat` in Appendix D and the full results of every domain in Appendix E.

Across languages. TIDE leads in every non-English language on both surrogates. On `qwen3.5-4b`, TIDE averages 0.970 against 0.892 for the best baseline, and its margin over the best baseline in each language ranges from 0.044 to 0.088. Chinese is the hardest language for TIDE, where it still achieves 0.941 against 0.858 for LRR, and German is the easiest, where it achieves 0.986 against 0.925 for Entropy. On `llama2-7b-chat`, TIDE averages 0.925 against 0.877 for DNA-DetectLLM. Chinese is again the hardest language, where TIDE also has its largest margin, 0.890 against 0.796 for DNA-DetectLLM, and Spanish is the easiest, where it achieves 0.952 against 0.921.

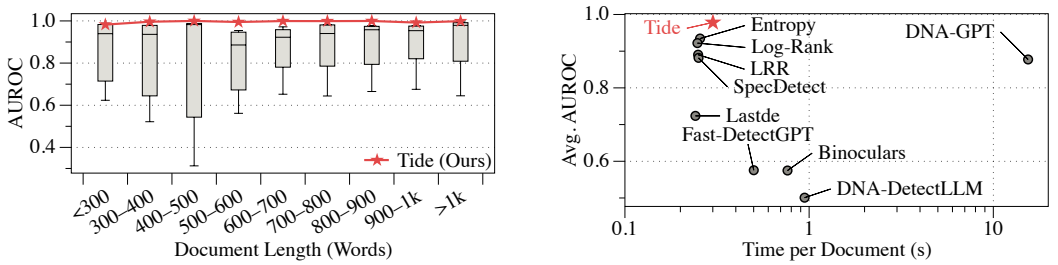
Stability across surrogates. A zero-shot detector can use any surrogate, so its accuracy should not depend on which surrogate is used. No baseline keeps its rank across the two surrogates. Binoculars rises from 0.575 to 0.922 and DNA-DetectLLM from 0.501 to 0.930, while Lastde drops from 0.724 to 0.304 and becomes inverted. The best baseline also changes, from Entropy on `qwen3.5-4b` to DNA-DetectLLM on `llama2-7b-chat`. In contrast, TIDE changes only from 0.978 to 0.961 and leads on both surrogates.

Against adversarial attacks. TIDE leads on five of the six attacks on both surrogates (Figure 3(b)) and on their average, with 0.936 against 0.899 on `qwen3.5-4b` and 0.954 against 0.939 on `llama2-7b-chat`. The largest margin of TIDE is on DeepWordBug with `qwen3.5-4b`, 0.978 against 0.882 for Entropy, and on Chinese back-translation with `llama2-7b-chat`, 0.968 against 0.910 for Fast-DetectGPT. On both back-translation attacks, which rewrite the text through another language, TIDE achieves at least 0.954 on `qwen3.5-4b` and 0.968 on `llama2-7b-chat`. DIPPER paraphrasing is the most difficult attack for TIDE, where its AUROC is only 0.059 below the best baseline on `qwen3.5-4b` and 0.054 below on `llama2-7b-chat`, and TIDE still outperforms six of the nine baselines.



(a) Per-Domain Breakdown **(b) Per-Attack Breakdown** **(c) Overall ROC Curves**

Figure 3: **Per-domain breakdown, per-attack breakdown and ROC curves on English, on the default surrogate.** Boxes span the nine baselines from minimum to maximum.



(a) AUROC against Document Length **(b) Inference Time against Detection Quality**

Figure 4: **Per-length breakdown and inference cost, on the default surrogate.** Boxes span the nine baselines from minimum to maximum.

5.3 Detection at Low False-Positive Rates

In practice, many applications require a detector to operate at a low FPR to minimize false alarms. We therefore measure the TPR at fixed FPRs on English. Since no method is trained, we use all 2,500 documents per class of each cell, so the lowest measurable FPR is 0.04%. Figure 3(c) shows the ROC curves averaged over the 25 English cells on a logarithmic FPR axis, with the nine baselines summarized by their range, median and best member. The lead of Tide in this figure increases as the FPR decreases. On `qwen3.5-4b`, Tide detects 97.0% of LLM-generated text against 87.4% for Entropy at 5% FPR and 91.6% against 67.0% at 1% FPR. At 0.1% FPR, it detects 79.1%, nearly twice the 41.0% of the best baseline, SpecDetect. On `llama2-7b-chat`, Tide also leads at every FPR, with 71.8% against 62.8% for Fast-DetectGPT at 0.1% FPR. We present the ROC curves on `llama2-7b-chat` in Figure 5(c) and the TPR of every method at every FPR in Appendix F.

5.4 Efficiency Analysis

In this section, we evaluate the inference cost of every method. On `qwen3.5-4b`, Tide takes 0.370 seconds per document, only 0.098 seconds more than the fastest baseline, Lastde, while its average AUROC is 0.254 higher (Figure 4(b)). This time is 1.2× that of a single forward pass on `qwen3.5-4b` and 1.8× on `llama2-7b-chat`, less than 2× because both sequences share one batched call. The three two-model methods, Fast-DetectGPT, Binoculars and DNA-DetectLLM, take 1.9 to 3.7×, and DNA-GPT, which regenerates the text ten times, takes 60 to 85×. We present the results on `llama2-7b-chat` in Figure 6(b) and the times of all methods in Appendix G.

5.5 Ablation Studies

Document length. We further test whether the accuracy of Tide depends on document length. Figure 4(a) groups the English documents by word count, computes the AUROC within each domain and averages it over the domains in each bin. Tide stays between 0.97 and 1.00 in every bin on both surrogates. The baselines vary more. In the 500–600-word bin, their median drops to 0.89 on `qwen3.5-4b` and 0.84 on `llama2-7b-chat`, while Tide scores 0.99 and 0.97. Every bin

above 600 words contains only QA-forum posts. We present the results on `llama2-7b-chat` in Figure 6(a) and the same analysis within each domain in Appendix H.

Reversal controls. Reversal disrupts the text, so the reverse loss might separate the classes only because of the disruption. We test this by replacing the reverse-order sequence with nine other corruptions of the token sequence, such as random permutation, block shuffling and random deletion or substitution of tokens. Every corruption that does not reverse the order separates the classes worse than token reversal on both surrogates, even when it increases the loss more. On the other languages with `llama2-7b-chat`, for example, token reversal achieves 0.925, while the strongest control achieves 0.860. Shuffling 16-token blocks keeps most of the signal, whereas shuffling 256-token blocks loses most of it, which shows that the signal comes from the local order of tokens. We present all nine corruptions in Appendix I.

Score components. We also ablate each component of the directional asymmetry score. Removing the reverse loss leaves the forward loss, the feature that many existing detectors build on, and lowers the AUROC on English from 0.995 to 0.975 and on the other languages from 0.970 to 0.887 on `qwen3.5-4b`, and from 0.992 to 0.943 and from 0.925 to 0.797 on `llama2-7b-chat`. The difference and the ratio alone are also weaker than their product in five of the six combinations of surrogate and condition. We present each component per domain in Appendix J.

Surrogate breadth. To test whether the gain depends on the surrogate, we run TIDE with 16 causal LLMs from four families (GPT-2 (Radford et al., 2019), Llama (Touvron et al., 2023; Grattafiori et al., 2024), Gemma (Team et al., 2024; 2025) and Qwen3.5 (Qwen Team, 2026)), from 0.12B to 9.2B parameters and from 2019 to 2026. With every surrogate, the directional asymmetry score separates the classes better than the forward loss alone on English and on each of the five other languages. The gain is largest where the forward loss is weakest. On English, it decreases from 0.230 for `gpt2` to 0.009 for `qwen3.5-9b`, whose forward loss already achieves 0.984. Moreover, the forward loss of the four GPT-2 models is inverted on the other languages (0.34 to 0.39), while their directional asymmetry score achieves 0.62 to 0.65. Every surrogate with at least one billion parameters exceeds 0.9 AUROC on English, including `gpt2-xl` from 2019. Recency matters more than size, as `qwen3.5-0.8b` outperforms `gemma2-9b` by 0.078 on English. We present the results of all 16 surrogates in Appendix K.

LLM-edited text. A growing number of people use LLMs to polish their own writing, and some applications may need to know how much an LLM is involved in a text, beyond a binary decision. Since TIDE outputs a scalar score, we test whether this score also reflects the degree of LLM involvement. For every human-written document, each of the five LLMs polishes the text while keeping its content and length, which adds 200,000 LLM-edited documents to our dataset. TIDE separates edited from unedited human-written text best on both surrogates, with 0.923 AUROC against 0.886 for Entropy on `qwen3.5-4b` and 0.940 against 0.893 for Fast-DetectGPT on `llama2-7b-chat`. It also ranks human-written, LLM-edited and LLM-generated text in this order in all 25 cells on `qwen3.5-4b` and in 22 of the 25 on `llama2-7b-chat`. We present the full results in Appendix L.

6 CONCLUSION

Most existing zero-shot detectors of LLM-generated text measure how predictable a text is to a surrogate LLM that reads it in forward order, yet in this order text from frontier LLMs is often hard to separate from human writing. In this paper, we observe that reading a text in reverse order makes every text harder to predict, but LLM-generated text loses disproportionately more predictability than human writing, an effect we call *directional generation asymmetry*. Based on this observation, we build TIDE, which obtains the forward and reverse losses of a text in one batched call to a frozen surrogate and combines them into a directional asymmetry score. On our new dataset spanning five domains, five frontier LLMs, six languages and six adversarial attacks, TIDE outperforms nine state-of-the-art zero-shot baselines, improving the average AUROC from the best baseline's 0.934 to 0.978 and nearly doubling the true-positive rate at a 0.1% false-positive rate. We discuss the limitations of TIDE in Appendix M.

AI USE STATEMENT

Since this paper studies the detection of LLM-generated text, we used frontier LLMs to generate the LLM-generated and LLM-edited documents of our dataset (§5.1 and Appendix A), while we kept the human-written documents from prior datasets unchanged. We also used generative AI tools to draft and polish the text of the paper, and we revised and checked every section. In addition, we used generative AI tools to search for related work, and we read and verified every related paper and checked and added every citation manually. We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including its text, claims and artifacts.

ETHICS STATEMENT

Our dataset builds on human-written documents from prior studies, and all of them are released under licenses that permit research use. The LLM-generated and LLM-edited documents are produced from these human-written documents as described in Appendix A. We intend the dataset as a testbed for researchers and developers to develop and evaluate detectors for LLM-generated text. We do not recommend deploying a detector developed or evaluated on this dataset, including TIDE, as the sole basis for consequential decisions in real-world applications, such as judgments of plagiarism or academic misconduct.

REPRODUCIBILITY STATEMENT

We present the construction of our dataset, including the data sources, the generation pipeline and the composition of each domain, in §5.1 and Appendix A. §4 describes every stage of TIDE, and §5.1 describes the surrogates, the baselines and the metrics. We will release the code of TIDE together with the dataset upon acceptance.

REFERENCES

Anthropic. Claude (sonnet 4.5), 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5. Large language model developed by Anthropic.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 24814–24836, 2024.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 51844–51865, 2025.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 61(2): 228–239, 2024.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality, 2023.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), 2024.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12463–12492, 2024.

European Parliament and Council of the European Union. Artificial intelligence act. Official Journal of the European Union, 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

Leon Fröhling and Arkaitz Zubiaga. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443, 2021.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*, 2021.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *IEEE Symposium on Security and Privacy Workshops (SPW)*, pp. 50–56, 2018.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Black-box detection of language model watermarks. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 77539–77562, 2025.

Google DeepMind. Gemini 3.1 flash-lite, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/. Large language model developed by Google DeepMind.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. Biscope: Ai-generated text detection by checking memorization of preceding tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 104065–104090, 2024a.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Guangyu Shen, Kaiyuan Zhang, Shengwei An, Guanhong Tao, and Xiangyu Zhang. Profiler: Black-box ai-generated text origin detection via context-aware inference pattern analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 24903–24923, 2025b.

Rui Guo, Weibin Zeng, Fuzhang Wu, Yan Kong, Yanjun Wu, Weiming Dong, et al. Hld: Approximate hierarchical linguistic distribution modeling for llm-generated text detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:88320–88347, 2024b.

Hans WA Hanley and Zakir Durumeric. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *International AAAI Conference on Web and Social Media (ICWSM)*, volume 18, pp. 542–556, 2024.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *International Conference on Machine Learning (ICML)*, 2024.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2251–2265, 2024.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. Detree: Detecting human-ai collaborative texts via tree-structured hierarchical representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:15077–15095, 2023.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 45408–45436, 2024.

Jiahao Huo, Shuliang Liu, Bin Wang, Junyan Zhang, Yibo Yan, Aiwei Liu, Xuming Hu, and Mingxun Zhou. Pmark: Towards robust and distortion-free semantic-level watermarking with channel constraints. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 27100–27126, 2026.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1808–1822, 2020.

Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. M-rangedetector: Enhancing generalization in machine-generated text detection through multi-range attention masks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8971–8983, 2025.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 8018–8025, 2020.

Tom Kempton, Julia Rozanova, Parameswaran Kamalaruban, Maeve Madigan, Karolina Wresilo, Yoann Launay, David Sutton, and Stuart Burrell. Dmap: A distribution map for text. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 54359–54386, 2026.

Sunnie SY Kim, Wesley Hanwen Deng, Jennifer Wortman Vaughan, Buxin Su, Weijie Su, Alekh Agarwal, Sharon Li, Martin Jaggi, Daniel G Goldstein, Nihar B Shah, et al. Use and effects of llms in peer review: A randomized experiment and survey at icml 2026. *arXiv preprint arXiv:2609.19420*, 2026.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, pp. 17061–17084, 2023.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 21258–21266, 2024.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:27469–27500, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research (TMLR)*, 2024. ISSN 2835-8856.

Lucio La Cava and Andrea Tagarelli. Luminol-aidetect: Fast zero-shot machine-generated text detection based on perplexity under text shuffling. *arXiv preprint arXiv:2604.25860*, 2026.

Harsh Nishant Lalai, Aashish Anantha Ramakrishnan, Raj Sanjay Shah, and Dongwon Lee. From intentions to techniques: A comprehensive taxonomy and challenges in text watermarking for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6147–6160, 2025.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *Network and Distributed System Security (NDSS) Symposium*, 2019.

Yuanfan Li, Qi Zhou, and Zexuan Xie. Learning from dictionary: Enhancing robustness of machine-generated text detection in zero-shot language via adversarial training. In *International Conference on Learning Representations (ICLR)*, 2026.

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6(12), 2025.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *Artificial Intelligence Review*, 59(2):74, 2026.

Ailiang Lin, Zhuoyun Li, Yusong Wang, Keyu Mao, Kotaro Funakoshi, and Manabu Okumura. Reverseeol: Improving training-free text embeddings via text reversal in decoder-only llms. *arXiv preprint arXiv:2606.05858*, 2026.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 6499–6519, 2024a.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024b.

Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. In-context watermarks for large language models. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 53136–53162, 2026.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, et al. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 91005–91056, 2025.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. Specdetect: Simple, fast, and training-free detection of llm-generated text via spectral analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 32356–32364, 2026.

Xiaoliang Luo, Xinyi Xu, Michael Ramscar, and Bradley C Love. Probability consistency in large language models: Theoretical foundations meet empirical discrepancies. *arXiv preprint arXiv:2505.08739*, 2025.

Shixuan Ma and Quan Wang. Zero-shot detection of llm-generated text using token cohesiveness. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 17538–17553, 2024.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9960–9987, 2023.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6348–6368, 2024.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar: generative ai detection via rewriting. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 6013–6030, 2024.

Mistral AI. Mistral large 3. https://docs.mistral.ai/models/mistral-large-3-25-12, 2025. Large language model developed by Mistral AI.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, pp. 24950–24962, 2023.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 119–126, 2020.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Vassilis Papadopoulos, Jérémie Wenger, and Clément Hongler. Arrows of time for large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Markmywords: Analyzing and evaluating language model watermarks. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 68–91, 2025.

Gerrit Quaremba, Elizabeth Black, Denny Vrandecic, and Elena Simperl. TSM-bench: Detecting LLM-generated text in real-world wikipedia editing practices. In *International Conference on Learning Representations (ICLR)*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Yiliao Song, Zhenqiao Yuan, Shuhai Zhang, Zhen Fang, Jun Yu, and Feng Liu. Deep kernel relative test for machine-generated text detection. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 90459–90485, 2025.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, 2023.

Jingtao Sun and Zhanglong Lv. Zero-shot detection of llm-generated text via text reorder. *Neurocomputing*, 631:129829, 2025.

Yuxia Sun, Ran Zhang, Aoxiang Sun, Xu Li, Zitao Liu, and Jingcai Guo. D&r: Recovery-based ai-generated text detection via a single black-box llm call. In *International Conference on Learning Representations (ICLR)*, 2026.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

China Law Translate. Measures for labeling of AI-generated synthetic content. China Law Translate, 2025. URL `https://www.chinalawtranslate.com/en/ai-labeling/`.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 2001–2016, 2021.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1702–1717, 2024.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*, 2024a.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3964–3992, 2024b.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1369–1407, 2024c.

Chenwang Wu, Yiu ming Cheung, Shuhai Zhang, Bo Han, and Defu Lian. Beyond raw detection scores: Markov-informed calibration for boosting machine-generated text detection. In *International Conference on Learning Representations (ICLR)*, 2026a.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:100369–100401, 2024.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics (CL)*, 51(1):275–338, 2025a.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. Who wrote this? the key to zero-shot llm-generated text detection is gecscore. In *International Conference on Computational Linguistics (COLING)*, pp. 10275–10292, 2025b.

Yihan Wu, Ruibo Chen, Georgios Milis, and Heng Huang. An ensemble framework for unbiased language model watermarking. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 17899–17917, 2026b.

Yihan Wu, Xuehao Cui, Ruibo Chen, and Heng Huang. Analyzing and evaluating unbiased language model watermark. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 142418–142433, 2026c.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free llm-generated text detection by mining token probability sequences. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 19072–19098, 2025.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 48572–48597, 2024.

Peigen Ye, Huali Ren, Zhengdao Li, Anli Yan, Hongyang Yan, Shaowei Wang, and Jin Li. Securing large language models: A survey of watermarking and fingerprinting techniques. *ACM Computing Surveys*, 58(7):1–35, 2026.

Varun Yerram, Rahul Madhavan, Sravanti Addepalli, Arun Suggala, Karthikeyan Shanmugam, and Prateek Jain. Time-reversal provides unsupervised feedback to llms. *Advances in Neural Information Processing Systems*, 37:29777–29806, 2024.

Xunjian Yin, Sitao Cheng, Yuxi Xie, Xinyu Hu, Li Lin, Xinyi Wang, Liangming Pan, William Yang Wang, and Xiaojun Wan. Ledom: Reverse language model. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 21306–21326, 2026.

Sicheng Yu, Xu Yuanchen, Cunxiao Du, Yanying Zhou, Minghui Qiu, Qianru Sun, Hao Zhang, and Jiawei Wu. Reverse modeling in large language models. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 306–320, 2025.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an LLM? benchmarking AI text detection in peer review. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. Text fluoroscopy: Detecting llm-generated text through intrinsic features. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15838–15846, 2024.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-flip: An optimally stable and watermarkable decoder for llms. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 32468–32496, 2025.

Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, Ying Yang, Shakeel Gavioli-Akilagun, and Chengchun Shi. Adadetectgpt: Adaptive detection of llm-generated text with statistical guarantees. *Advances in Neural Information Processing Systems*, 38:89077–89118, 2025.

Hongyi Zhou, Jin Zhu, Kai Ye, Ying Yang, Erhan Xu, and Chengchun Shi. Learn-to-distance: Distance learning for detecting LLM-generated text. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:1745–1768, 2025.

APPENDIX

The appendix is organized as follows.

- Appendix A: Construction details of our dataset, including the data sources, the generation pipeline, the translations, the adversarial attacks and the word counts of each domain.
- Appendix B: Results of all variants of Fast-DetectGPT, Lastde and SpecDetect on both surrogates.
- Appendix C: Paired significance tests of TIDE against each baseline on all languages and attacks.
- Appendix D: Per-domain, per-attack, ROC, per-length and inference-cost results on the second surrogate, `llama2-7b-chat`.
- Appendix E: Detailed AUROC of every method in each domain on both surrogates.
- Appendix F: True-positive rates of every method at fixed false-positive rates of 10%, 5%, 1% and 0.1%.
- Appendix G: Wall-clock inference time of every method and its ratio to a single forward pass.
- Appendix H: Detection performance across document lengths within each domain.
- Appendix I: Comparison of token reversal with nine other corruptions of the token sequence.
- Appendix J: Per-domain ablation of each component of the directional asymmetry score.
- Appendix K: Detection performance of TIDE with 16 surrogate LLMs from four model families.
- Appendix L: Detection results on LLM-edited text, including the score distributions of human-written, LLM-edited and LLM-generated text.
- Appendix M: Limitations of TIDE and directions for future work.

A DATASET CONSTRUCTION AND STATISTICS

This section presents the details of the dataset in §5.1.

Sources. The human-written documents come from existing detection benchmarks, including Yelp reviews and XSum news articles from DetectRL (Wu et al., 2024), OpenReview peer reviews from Yu et al. (2026), Reddit ELI5 answers from M4 (Wang et al., 2024c) and ASAP and TOEFL11 student essays from OUTFOX (Koike et al., 2024). We keep $2{,}500$ per domain after de-duplication.

Generation. Each LLM-generated document is written on the topic of one human-written document, and the prompt asks the LLM to match its word count. The LLM never sees the human-written text while it writes. For QA-forum posts and essays, the LLM receives the question or the essay prompt. For the other three domains, a first call condenses the human-written document into a structural template (peer reviews), a summary (news) or a keyword list (customer reviews), and a second call writes the new document from it. Each generated peer review is written for the reviewed paper itself and follows the template and the rating of the human reviewer. Every prompt asks for plain prose without headings or markup, and we remove generations that refuse the task or miss the word count. Table 2 shows that the word counts of the three classes are close in every domain. For the LLM-edited documents, each LLM polishes every human-written document while preserving its content and length. We present the results on LLM-edited documents in Appendix L.

Languages and attacks. The five other languages are translations of the English test documents of both classes by the Google Translate API, so each language cell pairs translated human-written text with translated LLM-generated text. The six attacks modify the LLM-generated test documents. DIPPER (Krishna et al., 2023) paraphrases them with a lexical and order diversity of 40, and two attacks translate them into Chinese or German and back (Macko et al., 2024). DeepWordBug (Gao et al., 2018), TextFooler (Jin et al., 2020) and TextBugger (Li et al., 2019) from TextAttack (Morris et al., 2020) perturb their characters and words, against the RADAR detector (Hu et al., 2023).

Table 2: Word counts of English documents per domain. P5 and P95 are the 5th and 95th percentiles.

	Human-written			LLM-edited			LLM-generated		
Domain	Mean	P5	P95	Mean	P5	P95	Mean	P5	P95
Customer reviews	232	167	331	216	153	312	231	163	332
News	241	165	312	238	163	314	253	173	331
Peer reviews	428	311	571	456	316	637	580	378	959
QA forum	412	196	860	397	184	843	429	191	948
Student essays	361	316	409	343	274	408	382	320	465

B Baseline Variants

Three baselines have two variants each, and §5.1 reports one variant of each, the same on both surrogates. Fast-DetectGPT runs either in the two-model setting of the main text or in a same-model setting, where the surrogate also supplies the sampling distribution. Lastde and SpecDetect each have an enhanced ++ variant. Table 3 reports every variant on both surrogates. On `qwen3.5-4b`, both ++ variants score below their base versions, and on `llama2-7b-chat`, the same-model Fast-DetectGPT, both Lastde variants and SpecDetect++ are inverted. For the base variant of SpecDetect, the released code applies the Fourier transform along an axis of length one, which leaves the sequence unchanged, so we apply it along the token positions, as the enhanced variant in the same code does. The TPR table in Appendix F also includes every variant.

Table 3: AUROC of every baseline variant by condition on the two surrogates. Marks are as in Table 1. Fast-DetectGPT$^{\circ}$ is the same-model setting, and ++ marks the enhanced variant.

	`qwen3.5-4b`				`llama2-7b-chat`			
Method	English	Languages	Attacks	Avg.	English	Languages	Attacks	Avg.
Entropy	0.984	0.892	0.899	0.934	0.905	0.806	0.716	0.846
Log-Rank	0.973	0.884	0.854	0.922	0.933	0.798	0.794	0.865
LRR	0.956	0.845	0.792	0.890	0.861	0.756	0.653	0.798
Fast-DetectGPT	0.649	0.500	0.588	0.576	0.976	0.859	0.939	0.926
Fast-DetectGPT$^{\circ}$	0.805	0.591	0.588	0.688	0.375	0.281	0.573	0.357
Binoculars	0.653	0.494	0.587	0.575	0.975	0.852	0.934	0.922
DNA-DetectLLM	0.512	0.475	0.572	0.501	0.974	0.877	0.922	0.930
Lastde	0.904	0.552	0.679	0.724	0.270	0.322	0.401	0.304
Lastde++	0.808	0.603	0.576	0.694	0.366	0.257	0.576	0.344
SpecDetect	0.966	0.834	0.702	0.882	0.895	0.624	0.705	0.767
SpecDetect++	0.825	0.630	0.507	0.708	0.398	0.275	0.563	0.365
DNA-GPT	0.927	0.844	0.796	0.877	0.787	0.591	0.703	0.700
TIDE (ours)	**0.995**	**0.970**	**0.936**	**0.978**	**0.992**	**0.925**	**0.954**	**0.961**

C Statistical Significance

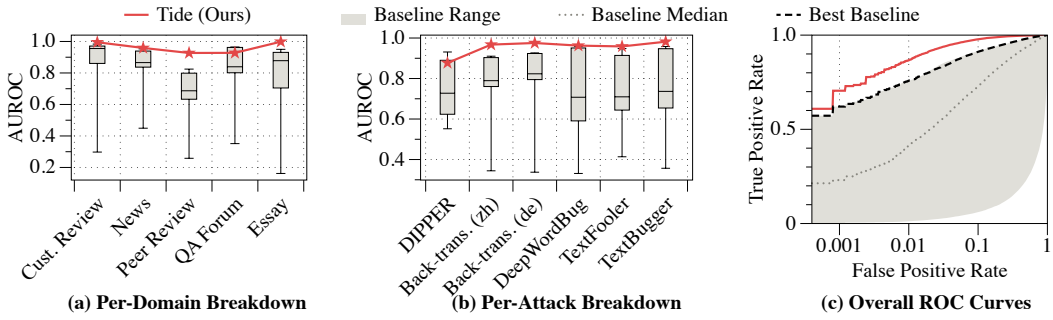
Since each margin in Table 1 averages over only 25 cells, we test whether it is significant. Table 4 presents paired tests of the margins in §5.2. Each test compares TIDE with one baseline on the same cells under one condition, such as English, one language or one attack, so both methods are scored on the same 500 human-written and 500 LLM-generated documents. We use the Wilcoxon signed-rank test and correct its p-values with the Holm method across the nine baselines within each condition. On `qwen3.5-4b`, every margin on English, on the other languages and on the attacks is significant at $p < 0.001$. On `llama2-7b-chat`, every margin on the other languages and on the attacks is significant at $p < 0.01$, and every margin on English is significant except those over the three two-model methods ($p = 0.24$).

Table 4: Paired tests of TIDE against each baseline. Δ is the mean AUROC difference, Wins counts the cells where TIDE scores higher and p is the Holm-corrected Wilcoxon p-value. English has 25 cells, the attacks 150 and the languages 125 (100 on `llama2-7b-chat`).

	English			Languages			Attacks		
Baseline	Δ	Wins	p	Δ	Wins	p	Δ	Wins	p
Surrogate: `qwen3.5-4b`									
Entropy	+0.012	22	<0.001	+0.078	116	<0.001	+0.036	111	<0.001
Log-Rank	+0.022	23	<0.001	+0.086	121	<0.001	+0.081	125	<0.001
LRR	+0.040	23	<0.001	+0.125	117	<0.001	+0.143	131	<0.001
Fast-DetectGPT	+0.347	25	<0.001	+0.470	125	<0.001	+0.348	141	<0.001
Binoculars	+0.342	25	<0.001	+0.476	125	<0.001	+0.349	141	<0.001
DNA-DetectLLM	+0.483	25	<0.001	+0.494	125	<0.001	+0.364	138	<0.001
Lastde	+0.091	25	<0.001	+0.418	125	<0.001	+0.257	138	<0.001
SpecDetect	+0.030	25	<0.001	+0.136	124	<0.001	+0.234	133	<0.001
DNA-GPT	+0.068	25	<0.001	+0.126	125	<0.001	+0.139	144	<0.001
Surrogate: `llama2-7b-chat`									
Entropy	+0.087	24	<0.001	+0.119	92	<0.001	+0.238	150	<0.001
Log-Rank	+0.059	23	<0.001	+0.127	94	<0.001	+0.160	144	<0.001
LRR	+0.131	25	<0.001	+0.169	93	<0.001	+0.301	149	<0.001
Fast-DetectGPT	+0.016	17	0.241	+0.065	72	<0.001	+0.015	98	0.006
Binoculars	+0.017	18	0.241	+0.073	75	<0.001	+0.020	102	0.004
DNA-DetectLLM	+0.018	18	0.241	+0.048	64	0.006	+0.032	105	<0.001
Lastde	+0.722	25	<0.001	+0.603	99	<0.001	+0.553	145	<0.001
SpecDetect	+0.097	25	<0.001	+0.301	100	<0.001	+0.249	136	<0.001
DNA-GPT	+0.205	25	<0.001	+0.334	96	<0.001	+0.251	150	<0.001

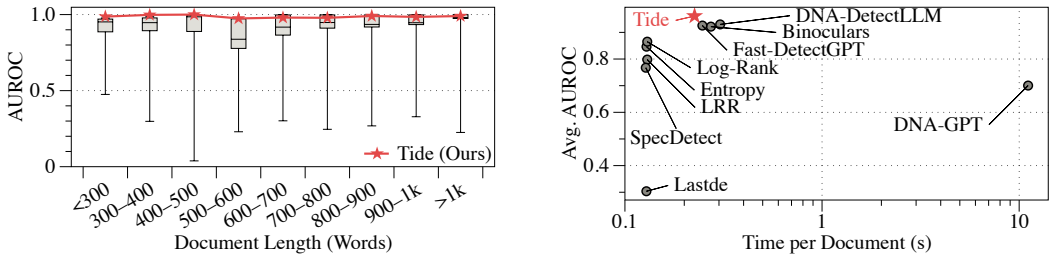
D RESULTS ON THE SECOND SURROGATE

Figure 5 and Figure 6 present the results of Figure 3 and Figure 4 on `llama2-7b-chat`. TIDE leads in four of the five domains, with its largest margin on peer reviews, and leads on five of the six attacks. The ROC curve of TIDE stays above the range of the baselines at low FPRs. Across document lengths, TIDE stays between 0.97 and 1.00 in every bin, and Fast-DetectGPT and Binoculars score slightly higher from 500 words up, where most documents are QA-forum posts. TIDE takes 1.8× the time of a single forward pass.



(a) Per-Domain Breakdown **(b) Per-Attack Breakdown** **(c) Overall ROC Curves**

Figure 5: **Per-domain breakdown, per-attack breakdown and ROC curves on English, on `llama2-7b-chat`.** Boxes span the nine baselines from minimum to maximum.



(a) AUROC against Document Length **(b) Inference Time against Detection Quality**

Figure 6: **Per-length breakdown and inference cost, on `llama2-7b-chat`.** Boxes span the nine baselines from minimum to maximum.

E Per-Domain Results

Since Table 1 averages over the five domains, it can hide a domain where a baseline is stronger. Table 5 and Table 6 therefore present the results of every method in each domain for the two surrogates. On `qwen3.5-4b`, TIDE leads in every domain, by 0.008 on customer reviews and news and by 0.043 to 0.089 on the other three domains. On `llama2-7b-chat`, TIDE leads in four of the five domains, and DNA-DetectLLM leads on QA-forum posts. The results of individual cells vary more than these averages.

F Additional Results on Detection at Low False-Positive Rates

Table 7 presents the TPR of every method variant at each FPR of §5.3. In each cell, the threshold for a target FPR is a quantile of the scores of the 2,500 human-written documents, and the TPR is measured on the 2,500 LLM-generated documents. With 2,500 human-written documents, the FPR changes in steps of 0.04%, so the 0.1% operating point in fact allows 0.12%, or 3 of the 2,500 documents. TIDE has the highest TPR at every FPR on both surrogates.

G Inference Cost

Table 8 presents the time per document of every method in §5.4. Each method recorded its wall-clock time per document while scoring the dataset, with eight scoring jobs running in parallel on eight RTX A6000 GPUs. On English, each method runs 30 scoring jobs, one for the human-written documents and one for each of the five LLMs in each of the five domains. Since the jobs differ in document length, we divide the time of each method by that of Entropy, a single forward pass, within each job and report the median ratio over the 30 jobs. Log-Rank, LRR, Lastde and SpecDetect, which also make a single forward pass, stay within 2% of Entropy on both surrogates. TIDE takes 1.21× the time of Entropy on `qwen3.5-4b` and 1.80× on `llama2-7b-chat`, since both sequences go through one batched call to one model. The three two-model methods make two forward passes through two models and take 1.9 to 3.7×, and DNA-GPT, which generates ten continuations per document, takes 60 to 85×. The table reports time only, and the two-model methods also hold two models in memory.

Table 5: AUROC of every method in each domain on `qwen3.5-4b`. Marks are as in Table 1.

	English, per generator					Other languages						
Method	Claude	DeepSeek	Gemini	GPT	Mistral	zh	de	ar	es	fr	Attacks	Avg.
Customer reviews												
Entropy	**1.000**	0.999	**1.000**	0.996	**1.000**	0.931	0.994	0.976	0.995	0.986	**0.981**	0.987
Log-Rank	**1.000**	0.999	**1.000**	0.990	**1.000**	0.934	0.984	0.964	0.993	0.978	0.964	0.982
LRR	0.999	0.981	0.997	0.976	0.999	0.913	0.912	0.923	0.981	0.949	0.918	0.959
Fast-DetectGPT	0.784	0.703	0.851	0.577	0.889	0.545	0.463	0.535	0.534	0.444	0.679	0.637
Binoculars	0.800	0.723	0.869	0.575	0.911	0.533	0.465	0.538	0.519	0.429	0.685	0.641
DNA-DetectLLM	0.694	0.625	0.821	0.537	0.834	0.492	0.453	0.576	0.498	0.409	0.683	0.602
Lastde	0.991	0.898	0.986	0.922	0.990	0.760	0.579	0.576	0.781	0.717	0.830	0.821
SpecDetect	0.999	0.988	0.999	0.981	**1.000**	0.860	0.939	0.869	0.962	0.948	0.869	0.947
DNA-GPT	0.988	0.970	0.985	0.878	0.995	0.819	0.952	0.896	0.936	0.916	0.896	0.930
TIDE (ours)	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**	**0.981**	**0.999**	**0.994**	**0.997**	**0.999**	0.972	**0.995**
Δ vs. best baseline	0.000	+0.001	0.000	+0.003	0.000	+0.047	+0.005	+0.018	+0.002	+0.013	-0.009	+0.008
News												
Entropy	**0.998**	0.994	**0.996**	**0.961**	0.997	0.694	0.961	0.922	**0.958**	**0.969**	**0.929**	0.943
Log-Rank	**0.998**	0.990	0.993	0.920	0.998	0.724	0.940	0.893	0.939	0.939	0.889	0.929
LRR	0.997	0.981	0.985	0.920	0.994	0.821	0.876	0.852	0.939	0.929	0.867	0.924
Fast-DetectGPT	0.732	0.360	0.688	0.538	0.816	0.564	0.593	0.557	0.518	0.480	0.587	0.585
Binoculars	0.739	0.349	0.692	0.527	0.829	0.545	0.581	0.542	0.507	0.464	0.584	0.578
DNA-DetectLLM	0.584	0.196	0.588	0.499	0.677	0.536	0.613	0.592	0.506	0.449	0.549	0.526
Lastde	0.966	0.878	0.927	0.822	0.982	0.720	0.423	0.384	0.690	0.574	0.756	0.738
SpecDetect	0.993	0.978	0.989	0.912	0.995	0.795	0.911	0.811	0.919	0.913	0.771	0.908
DNA-GPT	0.937	0.908	0.891	0.693	0.958	0.648	0.879	0.422	0.830	0.887	0.772	0.802
TIDE (ours)	0.994	**0.998**	**0.996**	0.913	**0.999**	**0.868**	**0.973**	**0.949**	0.956	0.947	0.873	**0.951**
Δ vs. best baseline	-0.004	+0.004	0.000	-0.048	+0.001	+0.047	+0.012	+0.027	-0.002	-0.022	-0.056	+0.008
Peer reviews												
Entropy	0.974	0.995	0.977	0.873	0.996	0.784	0.874	0.773	0.888	0.842	0.785	0.887
Log-Rank	0.978	0.994	0.978	0.707	0.995	0.782	0.826	0.727	0.825	0.831	0.697	0.849
LRR	0.943	0.949	0.940	0.590	0.979	0.829	0.694	0.625	0.751	0.717	0.580	0.782
Fast-DetectGPT	0.518	0.378	0.851	0.288	0.599	0.470	0.307	0.312	0.455	0.424	0.462	0.460
Binoculars	0.506	0.369	0.855	0.288	0.590	0.472	0.302	0.309	0.452	0.417	0.459	0.456
DNA-DetectLLM	0.269	0.142	0.766	0.303	0.320	0.500	0.286	0.311	0.395	0.371	0.511	0.379
Lastde	0.893	0.817	0.910	0.429	0.912	0.475	0.339	0.228	0.414	0.461	0.394	0.570
SpecDetect	0.954	0.980	0.973	0.615	0.990	0.686	0.697	0.600	0.749	0.715	0.430	0.763
DNA-GPT	0.907	0.964	0.906	0.726	0.985	0.757	0.824	0.828	0.865	0.850	0.703	0.847
TIDE (ours)	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**	**0.913**	**0.993**	**0.928**	**0.986**	**0.980**	**0.941**	**0.976**
Δ vs. best baseline	+0.022	+0.005	+0.022	+0.126	+0.004	+0.084	+0.119	+0.100	+0.098	+0.130	+0.156	+0.089
QA forum												
Entropy	0.982	0.990	0.969	0.929	0.993	0.767	0.855	0.798	0.875	0.866	0.853	0.898
Log-Rank	0.990	0.987	0.985	0.918	0.993	0.820	0.851	0.814	0.885	0.880	0.826	0.904
LRR	0.986	0.965	0.972	0.906	0.987	0.847	0.825	0.793	0.877	0.854	0.796	0.892
Fast-DetectGPT	0.841	0.470	0.983	0.837	0.774	0.597	0.525	0.664	0.603	0.619	0.684	0.691
Binoculars	0.859	0.469	0.988	0.846	0.786	0.600	0.529	0.677	0.605	0.618	0.687	0.697
DNA-DetectLLM	0.627	0.294	0.970	0.804	0.555	0.558	0.556	0.732	0.602	0.605	0.711	0.638
Lastde	0.981	0.885	0.967	0.897	0.968	0.760	0.584	0.548	0.698	0.706	0.691	0.790
SpecDetect	0.991	0.974	0.990	0.946	0.992	0.846	0.814	0.801	0.866	0.858	0.662	0.886
DNA-GPT	0.967	0.956	0.962	0.795	0.983	0.797	0.877	0.846	0.885	0.886	0.742	0.881
TIDE (ours)	**0.998**	**1.000**	**0.998**	**0.991**	**1.000**	**0.945**	**0.962**	**0.944**	**0.976**	**0.957**	**0.897**	**0.970**
Δ vs. best baseline	+0.007	+0.010	+0.008	+0.045	+0.007	+0.098	+0.085	+0.098	+0.091	+0.071	+0.044	+0.066
Student essays												
Entropy	**1.000**	**1.000**	0.999	0.975	**1.000**	0.808	0.941	0.908	0.982	0.960	0.948	0.956
Log-Rank	0.999	0.997	0.995	0.932	**1.000**	0.889	0.894	0.889	0.967	0.935	0.896	0.945
LRR	0.995	0.985	0.970	0.897	0.997	0.877	0.736	0.794	0.941	0.863	0.801	0.896
Fast-DetectGPT	0.665	0.432	0.769	0.147	0.732	0.589	0.346	0.507	0.435	0.409	0.527	0.505
Binoculars	0.666	0.429	0.779	0.139	0.744	0.582	0.344	0.486	0.429	0.404	0.521	0.502
DNA-DetectLLM	0.389	0.180	0.630	0.064	0.433	0.375	0.322	0.465	0.365	0.319	0.403	0.359
Lastde	0.983	0.918	0.945	0.751	0.987	0.828	0.207	0.304	0.522	0.526	0.722	0.699
SpecDetect	0.999	0.994	0.995	0.910	0.999	0.938	0.739	0.828	0.907	0.881	0.777	0.906
DNA-GPT	0.993	0.995	0.984	0.858	0.999	0.855	0.894	0.908	0.943	0.898	0.868	0.927
TIDE (ours)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.997**	**1.000**	**0.999**	**1.000**	**1.000**	**0.995**	**0.999**
Δ vs. best baseline	0.000	0.000	+0.001	+0.025	0.000	+0.059	+0.059	+0.091	+0.018	+0.040	+0.047	+0.043

Table 6: AUROC of every method in each domain on `llama2-7b-chat`. Marks are as in Table 1.

	English, per generator					Other languages						
Method	Claude	DeepSeek	Gemini	GPT	Mistral	zh	de	ar	es	fr	Attacks	Avg.
Customer reviews												
Entropy	**1.000**	0.992	0.993	0.959	0.999	0.945	0.972	—	0.991	0.964	0.896	0.971
Log-Rank	**1.000**	0.999	0.999	0.987	**1.000**	0.935	0.973	—	0.987	0.957	0.958	0.980
LRR	0.999	0.944	0.983	0.948	0.995	0.931	0.910	—	0.978	0.927	0.867	0.948
Fast-DetectGPT	0.967	0.986	0.967	0.998	0.994	0.906	0.904	—	0.946	0.920	0.963	0.955
Binoculars	0.977	0.990	0.977	**0.999**	0.996	0.904	0.915	—	0.955	0.928	0.969	0.961
DNA-DetectLLM	0.993	0.994	0.990	**0.999**	0.998	0.954	0.968	—	0.987	0.971	0.978	0.983
Lastde	0.314	0.177	0.450	0.424	0.493	0.236	0.101	—	0.104	0.156	0.521	0.298
SpecDetect	0.983	0.987	0.989	0.934	0.999	0.710	0.734	—	0.663	0.727	0.881	0.861
DNA-GPT	0.896	0.938	0.802	0.846	0.951	0.320	0.661	—	0.639	0.573	0.833	0.746
TIDE (ours)	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**	**0.973**	**0.996**	—	**0.995**	**0.996**	**0.989**	**0.995**
△ vs. best baseline	0.000	+0.001	+0.001	0.000	0.000	+0.019	+0.023	—	+0.004	+0.025	+0.011	+0.012
News												
Entropy	0.960	0.880	0.768	0.648	0.955	0.873	0.917	—	0.946	0.912	0.594	0.845
Log-Rank	0.992	0.986	0.935	0.848	0.997	0.847	0.926	—	0.923	0.917	0.808	0.918
LRR	0.967	0.849	0.797	0.770	0.950	0.894	0.905	—	**0.953**	0.918	0.655	0.866
Fast-DetectGPT	0.968	0.985	0.891	0.980	0.986	0.878	0.959	—	0.900	0.934	0.918	0.940
Binoculars	0.977	0.986	0.900	**0.981**	0.990	0.875	0.952	—	0.897	0.914	0.919	0.939
DNA-DetectLLM	0.980	0.981	0.900	0.980	0.991	**0.928**	**0.975**	—	0.951	**0.948**	0.896	0.953
Lastde	0.537	0.443	0.425	0.499	0.604	0.343	0.158	—	0.539	0.369	0.576	0.449
SpecDetect	0.978	0.978	0.929	0.857	0.992	0.679	0.761	—	0.628	0.766	0.805	0.837
DNA-GPT	0.776	0.793	0.530	0.481	0.769	0.169	0.548	—	0.373	0.528	0.605	0.557
TIDE (ours)	**0.993**	**1.000**	**0.992**	0.950	**1.000**	0.872	0.958	—	0.952	0.939	**0.934**	**0.959**
△ vs. best baseline	+0.001	+0.014	+0.057	-0.031	+0.003	-0.056	-0.017	—	-0.001	-0.009	+0.015	+0.006
Peer reviews												
Entropy	0.846	0.929	0.830	0.291	0.965	0.475	0.607	—	0.617	0.584	0.477	0.662
Log-Rank	0.877	0.958	0.878	0.325	0.984	0.507	0.573	—	0.630	0.581	0.555	0.687
LRR	0.791	0.703	0.726	0.185	0.903	0.462	0.457	—	0.442	0.449	0.362	0.548
Fast-DetectGPT	**0.993**	**1.000**	0.996	0.730	0.999	0.479	0.645	—	0.766	0.776	0.859	0.825
Binoculars	0.989	0.999	0.995	0.672	0.999	0.479	0.645	—	0.762	0.773	0.836	0.815
DNA-DetectLLM	0.987	0.999	0.993	0.589	0.999	0.462	0.653	—	0.755	0.767	0.783	0.799
Lastde	0.158	0.030	0.184	0.212	0.044	0.588	0.330	—	0.390	0.381	0.265	0.258
SpecDetect	0.693	0.867	0.809	0.505	0.933	0.519	0.473	—	0.528	0.520	0.487	0.633
DNA-GPT	0.832	0.951	0.777	0.701	0.972	0.701	0.715	—	0.767	0.689	0.759	0.786
TIDE (ours)	0.992	0.999	**0.999**	**0.973**	**1.000**	**0.801**	**0.826**	—	**0.902**	**0.834**	**0.944**	**0.927**
△ vs. best baseline	-0.001	-0.001	+0.003	+0.243	+0.001	+0.100	+0.111	—	+0.135	+0.058	+0.085	+0.102
QA forum												
Entropy	0.991	0.982	0.935	0.770	0.991	0.687	0.736	—	0.815	0.801	0.686	0.839
Log-Rank	0.991	0.984	0.967	0.801	0.996	0.733	0.759	—	0.827	0.804	0.770	0.863
LRR	0.987	0.930	0.926	0.753	0.984	0.773	0.711	—	0.789	0.764	0.676	0.829
Fast-DetectGPT	**1.000**	**1.000**	**1.000**	**0.993**	**1.000**	0.822	0.913	—	0.956	0.955	**0.986**	0.962
Binoculars	**1.000**	**1.000**	**1.000**	0.992	**1.000**	0.825	0.913	—	0.956	0.955	0.983	0.962
DNA-DetectLLM	**1.000**	**1.000**	**1.000**	0.989	**1.000**	**0.841**	**0.919**	—	**0.962**	**0.965**	0.970	**0.965**
Lastde	0.239	0.068	0.509	0.336	0.283	0.493	0.319	—	0.417	0.344	0.506	0.351
SpecDetect	0.951	0.916	0.971	0.787	0.991	0.680	0.664	—	0.692	0.671	0.688	0.801
DNA-GPT	0.850	0.796	0.732	0.714	0.885	0.688	0.692	—	0.713	0.741	0.666	0.748
TIDE (ours)	0.989	0.993	0.983	0.945	0.995	0.812	0.858	—	0.914	0.885	0.901	0.928
△ vs. best baseline	-0.011	-0.007	-0.017	-0.048	-0.005	-0.029	-0.061	—	-0.048	-0.080	-0.085	-0.037
Student essays												
Entropy	0.999	0.999	0.994	0.948	**1.000**	0.667	0.820	—	0.921	0.869	0.927	0.915
Log-Rank	0.995	0.991	0.974	0.872	0.999	0.652	0.741	—	0.867	0.809	0.880	0.878
LRR	0.973	0.927	0.881	0.657	0.988	0.682	0.631	—	0.820	0.721	0.706	0.799
Fast-DetectGPT	0.988	0.998	0.996	0.989	0.993	0.815	0.822	—	0.935	0.957	0.967	0.946
Binoculars	0.990	0.998	0.996	0.985	0.994	0.767	0.764	—	0.912	0.940	0.962	0.931
DNA-DetectLLM	0.998	**1.000**	0.999	0.993	0.999	0.793	0.827	—	0.949	0.962	0.981	0.950
Lastde	0.075	0.015	0.045	0.065	0.113	0.414	0.077	—	0.439	0.238	0.138	0.162
SpecDetect	0.913	0.915	0.815	0.691	0.982	0.517	0.440	—	0.528	0.582	0.662	0.705
DNA-GPT	0.776	0.789	0.663	0.636	0.823	0.420	0.589	—	0.672	0.630	0.653	0.665
TIDE (ours)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.992**	**0.997**	—	**0.998**	**0.998**	**1.000**	**0.998**
△ vs. best baseline	+0.001	0.000	+0.001	+0.007	0.000	+0.177	+0.170	—	+0.049	+0.036	+0.019	+0.048

Table 7: True-positive rate at fixed false-positive rates on English for every variant. Marks are as in Table 1. Fast-DetectGPT° is the same-model setting, and ++ the enhanced variant.

	qwen3.5-4b					llama2-7b-chat				
		TPR at FPR					TPR at FPR			
Method	AUROC	10%	5%	1%	0.1%	AUROC	10%	5%	1%	0.1%
Entropy	0.973	0.925	0.874	0.670	0.387	0.892	0.729	0.632	0.419	0.247
Log-Rank	0.959	0.896	0.837	0.662	0.353	0.920	0.803	0.714	0.480	0.239
LRR	0.936	0.835	0.749	0.548	0.327	0.835	0.598	0.476	0.280	0.150
Fast-DetectGPT	0.622	0.277	0.187	0.070	0.017	0.964	0.912	0.870	0.759	0.628
Fast-DetectGPT°	0.763	0.464	0.340	0.139	0.052	0.365	0.051	0.023	0.004	0.000
Binoculars	0.626	0.318	0.228	0.094	0.019	0.962	0.911	0.873	0.751	0.550
DNA-DetectLLM	0.506	0.207	0.141	0.059	0.009	0.962	0.916	0.878	0.748	0.545
Lastde	0.869	0.656	0.544	0.324	0.143	0.273	0.060	0.038	0.012	0.004
Lastde++	0.766	0.458	0.332	0.138	0.044	0.356	0.050	0.024	0.004	0.001
SpecDetect	0.949	0.868	0.798	0.617	0.410	0.868	0.650	0.543	0.353	0.160
SpecDetect++	0.784	0.481	0.350	0.156	0.040	0.387	0.059	0.025	0.004	0.000
DNA-GPT	0.878	0.677	0.556	0.315	0.117	0.765	0.305	0.189	0.050	0.004
TIDE (ours)	**0.993**	**0.982**	**0.970**	**0.916**	**0.791**	**0.990**	**0.978**	**0.957**	**0.869**	**0.718**
Δ vs. best baseline	+0.020	+0.057	+0.096	+0.246	+0.381	+0.026	+0.062	+0.079	+0.110	+0.090

Table 8: Wall-clock seconds per document (mean over cells) and the median per-cell ratio to Entropy, a single forward pass, on the two surrogates. Fast-DetectGPT° is the same-model setting, and ++ the enhanced variant.

	qwen3.5-4b		llama2-7b-chat	
Method	s/document	× Entropy	s/document	× Entropy
Entropy	0.277	1.00	0.145	1.00
Log-Rank	0.278	1.00	0.146	1.01
LRR	0.281	1.02	0.147	1.01
Fast-DetectGPT	0.555	1.98	0.275	1.90
Fast-DetectGPT°	0.295	1.06	0.150	1.03
Binoculars	0.814	2.97	0.303	2.12
DNA-DetectLLM	1.003	3.68	0.334	2.36
Lastde	0.272	0.98	0.144	1.00
Lastde++	0.398	1.44	0.242	1.71
SpecDetect	0.275	0.99	0.144	0.99
SpecDetect++	0.313	1.13	0.155	1.07
DNA-GPT	15.513	59.84	11.468	85.10
TIDE (ours)	0.370	1.21	0.262	1.80

H DOCUMENT LENGTH PER DOMAIN

In Figure 4(a), every bin above 600 words contains only QA-forum posts, so the effects of length and domain are mixed. Figure 7 and Figure 8 therefore repeat the length analysis of §5.5 within each domain. The bin edges are chosen per domain so that every bin contains documents, and the ticks mark the edges. Within a domain, we pool the five generators against the human-written documents, so a trend within a panel reflects length alone. TIDE stays above 0.98 in every bin of customer reviews, peer reviews and student essays and at least 0.95 in every bin of news and QA-forum posts on both surrogates. On news with `qwen3.5-4b` and on QA-forum posts with `llama2-7b-chat`, the strongest baseline scores slightly higher, as in Appendix E.

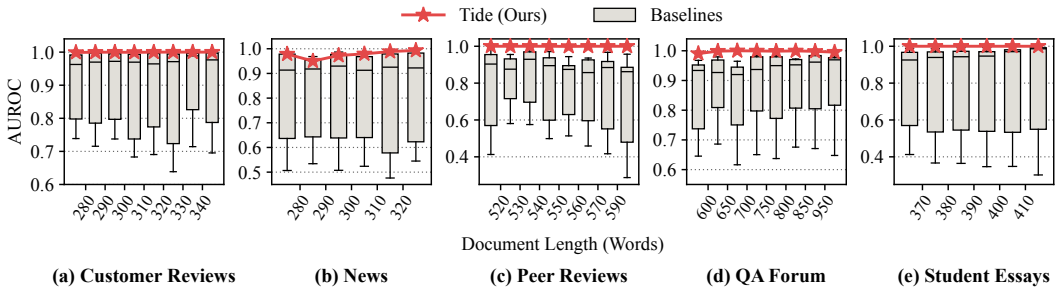


Figure 7: **Per-length breakdown within each domain, on `qwen3.5-4b`.** Boxes span the nine baselines from minimum to maximum.

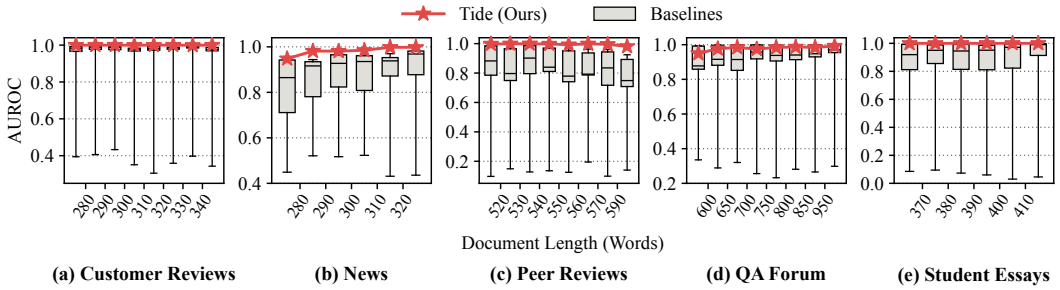


Figure 8: **Per-length breakdown within each domain, on `llama2-7b-chat`.** Boxes span the nine baselines from minimum to maximum.

I Additional Results on Reversal Controls

In §3, we attribute the increase from the forward loss to the reverse loss to the direction of reading. The reverse loss could instead measure only how far a scrambled sequence is from the text that the surrogate was trained on, and then any corruption of the sequence would work as well. Table 9 tests this for the reversal controls of §5.5 by recomputing the score of Equation 4 with the reverse-order sequence replaced by nine other corruptions of the forward-order sequence. The first is a random permutation of the tokens. Four of them shuffle blocks of 4, 16, 64 or 256 tokens and keep the order inside each block. Two reverse the order of the words or of the sentences and keep the order inside each unit. The last two delete or substitute 30% of the tokens and keep the order of the rest.

A corruption that increases the loss more may separate the classes better for that reason alone. We therefore compare each corruption at its inflation, the loss of the corrupted sequence divided by the forward loss, and fit a linear trend of AUROC against inflation over the seven corruptions that carry no direction, namely the shuffles, the permutation, the deletion and the substitution. The residual of token reversal above this trend measures what the direction adds beyond the size of the disruption.

Every corruption that carries no direction scores below token reversal on both surrogates, on English and on the other languages. On English, the strongest controls are close to token reversal, at most 0.002 below it on `qwen3.5-4b` and 0.010 below it on `llama2-7b-chat`, with residuals of 0.020 and 0.002. Word-order reversal, itself a reversal, scores slightly higher than token reversal on English with `qwen3.5-4b` (0.996 against 0.995). On the other languages, the margins are larger. Token reversal outperforms every control on both surrogates, with a lower inflation than random permutation and random substitution and with residuals of 0.045 on `qwen3.5-4b` and 0.072 on `llama2-7b-chat`. Every pairwise difference on the other languages is significant at $p < 10^{-8}$ by a Wilcoxon test over the 125 language cells of `qwen3.5-4b` and the 100 of `llama2-7b-chat`. The block shuffles show that the signal comes from the local order of tokens. Compared with token reversal, shuffling 16-token blocks loses 0.010 and 0.067 AUROC on the two surrogates, and shuffling 256-token blocks loses 0.254 and 0.284. Reversing the order of the words or of the sentences, which keeps every token next to its original neighbors on one side, loses more than

shuffling 16-token blocks. The word-order row includes Chinese, which has no spaces between words, so the operation leaves Chinese text unchanged. Without Chinese, word-order reversal scores 0.921 on `qwen3.5-4b` and 0.792 on `llama2-7b-chat`.

Table 9: AUROC of the directional asymmetry score when the reverse-order sequence is replaced by nine other corruptions of the forward-order sequence, on English and on the other languages, for the two surrogates. As in Table 1, the other languages exclude Arabic for `llama2-7b-chat`. Infl. (inflation) is the loss of the second sequence divided by the forward loss, and marks are as in Table 1.

	`qwen3.5-4b`				`llama2-7b-chat`			
	English		Languages		English		Languages	
Second sequence	AUROC	Infl.	AUROC	Infl.	AUROC	Infl.	AUROC	Infl.
Token reversal (TIDE)	0.995	2.71	**0.970**	2.82	**0.992**	3.21	**0.925**	3.36
4-token block shuffle	0.994	2.39	0.948	2.32	0.980	2.49	0.855	2.65
16-token block shuffle	0.993	1.60	0.960	1.54	0.967	1.59	0.858	1.64
64-token block shuffle	0.941	1.20	0.896	1.17	0.848	1.19	0.773	1.20
256-token block shuffle	0.716	1.04	0.715	1.03	0.641	1.04	0.641	1.04
Token permutation	0.989	3.31	0.924	3.29	0.982	3.62	0.848	3.89
Word-order reversal	**0.996**	2.42	0.864	1.94	0.951	2.35	0.716	1.78
Sentence-order reversal	0.977	1.16	0.890	1.09	0.813	1.11	0.682	1.06
30% token deletion	0.988	2.07	0.937	2.10	0.973	2.19	0.860	2.48
30% token substitution	0.986	3.64	0.910	3.47	0.959	3.40	0.825	3.52
Residual above the control trend	+0.020		+0.045		+0.002		+0.072	

J SCORE COMPONENTS PER DOMAIN

Table 10 presents the score components of §5.5 for each domain and condition, with each component of the directional asymmetry score evaluated on its own. Each entry averages the five generators and all variants of a condition within one domain, and Macro averages the five domains. The attack condition covers all six attacks, as the Attacks column of Table 1 does. Each component is read in a fixed direction chosen in advance, in which a lower forward loss and higher values of the other components indicate LLM-generated text, so a value below 0.5 means that the component is inverted. With D the difference and r the ratio, the score of Equation 4 expands as $D \cdot r = D + D^2/L_{\text{fwd}}$. The first term keeps the difference, and the second multiplies it by the relative difference D/L_{fwd}, so a large difference on a text that is already easy to predict scores highest. The D^2/L_{fwd} row evaluates this second term alone. On its own, the reverse loss is the weakest component on every macro average, and it carries the signal only together with the forward loss. The two factors are strongest in different domains. The ratio is the stronger factor on news in five of the six combinations of surrogate and condition, and the difference is the stronger factor on peer reviews in all six, where the forward loss separates the classes worst. On the macro average, the product outperforms both factors in five combinations and is 0.001 below the ratio in the sixth. The second term alone is within 0.005 of the product on every macro average, and no document in the dataset has $D < 0$, so the two scores behave alike in practice. We use the product because its two factors, the difference and the ratio, are easy to interpret (§4.4).

Table 10: AUROC of each component of the directional asymmetry score on its own, per domain and condition, on the two surrogates. $D = L_{\text{rev}} - L_{\text{fwd}}$ is the difference, $r = L_{\text{rev}}/L_{\text{fwd}}$ is the ratio and marks are as in Table 1.

	qwen3.5-4b						llama2-7b-chat					
Component	Reviews	News	Peer rev.	QA forum	Essays	Macro	Reviews	News	Peer rev.	QA forum	Essays	Macro
English												
L_{fwd}	0.998	0.977	0.936	0.975	0.988	0.975	0.997	0.957	0.836	0.948	0.977	0.943
L_{rev}	0.479	0.411	0.794	0.525	0.828	0.608	0.803	0.809	0.865	0.698	0.993	0.834
D	0.995	0.931	0.999	0.984	**1.000**	0.982	0.999	0.978	0.992	0.956	**1.000**	0.985
r	**1.000**	**0.986**	0.996	0.997	**1.000**	**0.996**	**1.000**	0.983	0.977	**0.983**	**1.000**	0.989
D^2/L_{fwd}	**1.000**	0.983	0.999	**0.998**	**1.000**	**0.996**	**1.000**	**0.987**	0.991	0.982	**1.000**	**0.992**
$D \cdot r$ (TIDE)	**1.000**	0.980	**1.000**	0.997	**1.000**	0.995	**1.000**	**0.987**	**0.993**	0.981	**1.000**	**0.992**
Other languages												
L_{fwd}	0.973	0.878	0.808	0.849	0.926	0.887	0.956	0.873	0.612	0.773	0.772	0.797
L_{rev}	0.673	0.606	0.714	0.594	0.882	0.694	0.813	0.643	0.729	0.644	0.929	0.752
D	0.989	0.905	0.949	0.921	**0.999**	0.953	0.980	0.885	0.832	0.816	0.994	0.902
r	0.993	0.933	0.934	0.941	0.997	0.960	0.988	0.927	0.791	0.864	0.980	0.910
D^2/L_{fwd}	**0.994**	**0.939**	0.957	0.956	**0.999**	0.969	**0.990**	**0.931**	0.834	**0.870**	0.995	0.924
$D \cdot r$ (TIDE)	**0.994**	**0.939**	**0.960**	**0.957**	**0.999**	**0.970**	**0.990**	0.930	**0.841**	0.867	**0.996**	**0.925**
All six attacks												
L_{fwd}	0.971	0.886	0.716	0.830	0.917	0.864	0.969	0.852	0.628	0.791	0.914	0.831
L_{rev}	0.464	0.427	0.839	0.561	0.821	0.622	0.768	0.783	0.898	0.716	0.986	0.830
D	0.934	0.790	**0.951**	0.855	0.992	0.904	0.978	0.924	**0.960**	0.881	**1.000**	0.949
r	**0.984**	**0.900**	0.897	0.898	0.992	0.934	**0.990**	0.919	0.881	0.887	0.999	0.935
D^2/L_{fwd}	0.977	0.882	0.932	**0.901**	**0.995**	**0.937**	**0.990**	0.933	0.933	**0.901**	**1.000**	0.951
$D \cdot r$ (TIDE)	0.972	0.873	0.941	0.897	**0.995**	0.936	0.989	**0.934**	0.944	**0.901**	**1.000**	**0.954**

K ADDITIONAL RESULTS ON SURROGATE BREADTH

Table 11: AUROC of TIDE with each of 16 causal LLMs as the surrogate. Size is in billions of parameters, Forward is the forward loss alone on English and Gain is the improvement of TIDE over it. The shaded row is the default surrogate. Languages excludes Arabic for `llama2-7b-chat`.

				English				
Family	Surrogate	Size	Year	Forward	TIDE	Gain	Languages	Attacks
GPT-2	`gpt2`	0.12	2019	0.634	0.864	+0.230	0.631	0.765
GPT-2	`gpt2-medium`	0.35	2019	0.693	0.896	+0.203	0.616	0.792
GPT-2	`gpt2-large`	0.77	2019	0.719	0.899	+0.180	0.633	0.795
GPT-2	`gpt2-xl`	1.50	2019	0.739	0.906	+0.167	0.647	0.799
Llama	`llama3.2-1b`	1.20	2024	0.887	0.981	+0.094	0.873	0.912
Llama	`llama3.2-3b`	3.20	2024	0.918	0.986	+0.068	0.912	0.912
Llama	`llama2-7b-chat`	6.70	2023	0.943	0.992	+0.049	0.925	0.954
Llama	`llama3.1-8b`	8.00	2024	0.925	0.976	+0.051	0.930	0.874
Gemma	`gemma3-270m`	0.27	2025	0.664	0.796	+0.132	0.760	0.692
Gemma	`gemma3-1b`	1.00	2025	0.839	0.972	+0.133	0.914	0.900
Gemma	`gemma2-2b`	2.60	2024	0.833	0.956	+0.123	0.833	0.858
Gemma	`gemma2-9b`	9.20	2024	0.860	0.912	+0.052	0.905	0.787
Qwen3.5	`qwen3.5-0.8b`	0.80	2026	0.921	0.990	+0.069	0.933	0.941
Qwen3.5	`qwen3.5-2b`	2.00	2026	0.960	0.997	+0.037	0.957	0.955
Qwen3.5	`qwen3.5-4b`	4.00	2026	0.975	0.995	+0.020	0.970	0.936
Qwen3.5	`qwen3.5-9b`	9.00	2026	0.984	0.993	+0.009	0.976	0.916

Table 11 presents the 16 surrogates of the surrogate breadth study in §5.5, each run with the same 1,024 input token limit as in §5.1. On English, TIDE achieves between 0.796 (`gemma3-270m`) and 0.997 (`qwen3.5-2b`), and even `gpt2-xl` from 2019 achieves 0.906. Recency matters more than size, as `qwen3.5-0.8b` outperforms `gemma2-9b` by 0.078 on English and by 0.028 on the

other languages. In three of the four families, the English AUROC peaks below the largest model, at `llama2-7b-chat`, `gemma3-1b` and `qwen3.5-2b`. The gain of TIDE over the forward loss on English decreases from 0.230 for `gpt2` to 0.009 for `qwen3.5-9b` as the forward loss approaches the ceiling, and it stays positive for all 16 surrogates. On each of the five other languages, the gain is also positive for all 16 surrogates, and it is largest for GPT-2 (0.25 to 0.29 on average). The forward loss of GPT-2 is inverted on the other languages (0.34 to 0.39), yet its directional asymmetry score achieves 0.62 to 0.65.

L ADDITIONAL RESULTS ON LLM-EDITED TEXT

LLMs are often used to polish human writing, and recent work detects text that an LLM has partly written or measures how much an LLM has edited it. For the LLM-edited text study of §5.5, the dataset pairs every LLM-generated document with an LLM-edited one, in which the same LLM polishes the human-written original while preserving its content and length. Since an edited document keeps human content in LLM wording, we evaluate it separately and leave it out of every other result. Table 12 reports the AUROC of every method on the three pairs of classes, and its Ordered column counts the cells in which a method ranks human-written, LLM-edited and LLM-generated text in this order. A value below 0.5 in the Edited/human column means that the method rates edited text as more human than the unedited original. Figure 9 and Figure 10 show the three score distributions for TIDE and the nine baselines, with scores min-max normalized within each (domain, generator) cell and pooled over the 25 English cells so that methods on different scales share one axis. Each violin is a kernel density estimate, and the cross marks the median (bar) and the interquartile range (line). TIDE separates edited from unedited human-written text best on both surrogates, with 0.923 against 0.886 for Entropy on `qwen3.5-4b` and 0.940 against 0.893 for Fast-DetectGPT on `llama2-7b-chat`. Which methods are inverted depends on the surrogate. On `qwen3.5-4b`, the three two-model methods score below 0.5, and on `llama2-7b-chat`, both Lastde variants and SpecDetect++ do, while the same-model Fast-DetectGPT is at chance.

Table 12: AUROC of every method on the three pairs of classes among human-written, LLM-edited and LLM-generated text on English, and the number of the 25 cells in which the method ranks the three classes in this order (Ordered). Column heads name the two classes compared, marks are as in Table 1 and Fast-DetectGPT° is the same-model setting.

	`qwen3.5-4b`				`llama2-7b-chat`			
Method	Edited/human	LLM/edited	LLM/human	Ordered	Edited/human	LLM/edited	LLM/human	Ordered
Entropy	<u>0.886</u>	<u>0.876</u>	<u>0.984</u>	24/25	0.769	0.780	0.905	23/25
Log-Rank	0.832	0.870	0.973	24/25	0.830	0.793	0.933	23/25
LRR	0.756	0.843	0.956	23/25	0.656	0.774	0.861	21/25
Fast-DetectGPT	0.491	0.663	0.649	11/25	<u>0.893</u>	0.832	<u>0.976</u>	24/25
Fast-DetectGPT°	0.618	0.730	0.805	19/25	0.500	0.364	0.375	3/25
Binoculars	0.482	0.672	0.653	12/25	0.891	<u>0.841</u>	0.975	24/25
DNA-DetectLLM	0.385	0.621	0.512	4/25	0.877	**0.860**	0.974	24/25
Lastde	0.706	0.780	0.904	21/25	0.323	0.427	0.270	0/25
Lastde++	0.612	0.737	0.808	19/25	0.475	0.384	0.366	3/25
SpecDetect	0.799	0.842	0.966	22/25	0.767	0.726	0.895	19/25
SpecDetect++	0.624	0.748	0.825	19/25	0.466	0.430	0.398	4/25
DNA-GPT	0.763	0.793	0.927	22/25	0.695	0.648	0.787	21/25
TIDE (ours)	**0.923**	**0.926**	**0.995**	25/25	**0.940**	0.822	**0.992**	22/25

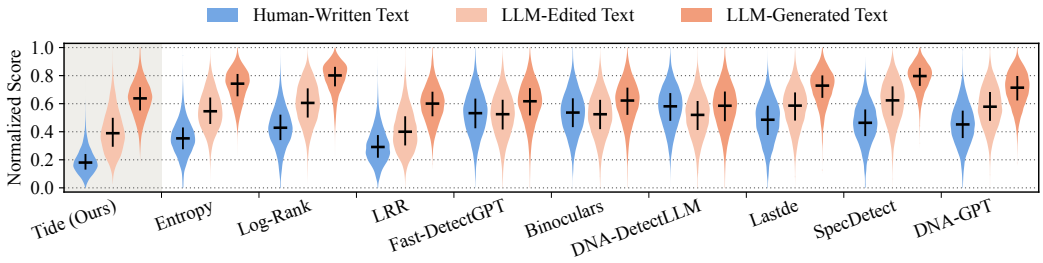


Figure 9: Scores of human-written, LLM-edited and LLM-generated text on `qwen3.5-4b`.

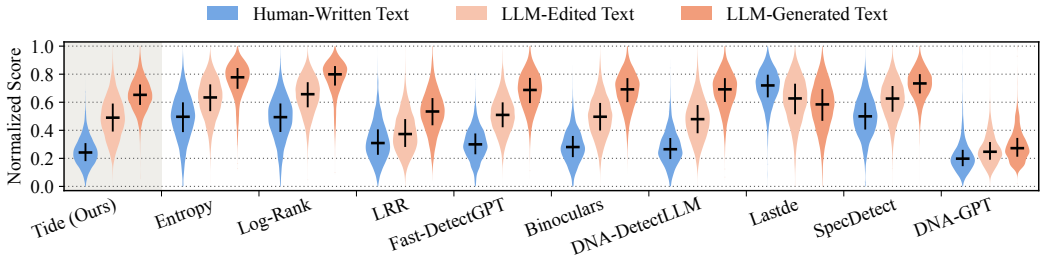


Figure 10: Scores of human-written, LLM-edited and LLM-generated text on `llama2-7b-chat`.

M LIMITATIONS

While we evaluate TIDE across diverse domains, generators, languages and adversarial attacks, our study has several limitations. First, our dataset cannot cover every domain, LLM, language, attack or document length that a detector may face in practice. For example, it contains few documents shorter than 150 words, such as tweets or abstracts. Second, like other zero-shot detectors, TIDE outputs one scalar score and thus targets the binary decision between human-written and LLM-generated text. Our study of LLM-edited text suggests that this score also reflects the degree of LLM involvement, but real-world use of LLMs in writing is more diverse, and finer-grained attribution of LLM involvement remains an open problem. Finally, a score from TIDE indicates likely LLM involvement and should not be the sole evidence of authorship, especially where a false positive can cause serious consequences.