AEyeDE: An Attention-Based Attribution Framework for AI-Generated Text Detection

Anonymous authors
Paper under double-blind review

Abstract

Detecting AI-generated text is becoming increasingly challenging as modern language models approach human-level fluency and can evade detectors that rely on surface statistics or likelihood-based signals. We propose AEyeDE, an attribution-driven approach to human-AI authorship detection that leverages model attention as a discriminative signal. Specifically, we extract attention-based attribution matrices for both human- and AI-generated text using a *proxy* Transformer model with white-box access and train a lightweight Convolutional Neural Network to learn representations from these attribution maps. Across encoder-decoder translation settings, our method consistently outperforms a text-only baseline. In decoder-only settings, it performs strongly in generator-specific detection, remains competitive on standard benchmarks, and shows robustness under cross-dataset transfer and alternative-spelling perturbations. We further show that attention maps exhibit recurring local structures whose relative frequencies differ consistently between human- and AI-generated text across datasets and proxy models. These findings suggest that attention-based attribution maps provide a complementary and interpretable signal for AI-generated text detection. We will make the code publicly available to support future research.

1 Introduction

The advent of large language models (LLMs) has enabled the generation of coherent, context-aware, and human-like text across a wide range of domains and languages (Naveed et al., 2023; Chang et al., 2024). While these advances unlock substantial benefits, they also raise critical challenges related to information integrity, authorship, and misuse, including large-scale misinformation in journalism and academic dishonesty in educational settings, where automated content generation threatens societal trust, originality, and assessment validity (Dugan et al., 2023; Wu et al., 2025; Liu et al., 2024b; Ali et al., 2025; Huang et al., 2025b; Bittle & El-Gayar, 2025).

In response, several AI-generated text detection methods have been proposed, including surface-statistical approaches exploiting cues such as perplexity, burstiness, and n-gram repetition (Gehrmann et al., 2019; Ippolito et al., 2020); likelihood-based methods that probe variation in a model's probability landscape via perturbation or re-sampling (Mitchell et al., 2023; Bao et al., 2023); supervised classifiers that fine-tune Transformer encoders on labeled data (Li et al., 2025; Zhu et al., 2025); and watermarking techniques that embed detectable signals for source attribution (Kirchenbauer et al., 2023; Liu et al., 2024a). However, each paradigm has intrinsic limitations: statistical and likelihood-based detectors degrade as LLMs are optimized to mimic human distributions through techniques such as RLHF (Christiano et al., 2017); supervised classifiers suffer under domain shift and unseen generators (Uchendu et al., 2021); and watermarking requires model-side cooperation and is fragile to paraphrasing, post-editing, or partial reuse (Liu et al., 2024a; Wang et al., 2025; Niess & Kern, 2025; Ahn et al., 2025). As a result, robust detection remains an open challenge, continually undermined by advances in generation quality and adversarial evasion strategies (Wu et al., 2025; 2024).

These challenges motivate a shift away from detecting *what* is written toward analyzing *how* text is produced. In particular, we hypothesize that token-token interaction structure provides a richer detection signal than

one-dimensional token-level statistics such as rank, likelihood, or perplexity. Text-only detectors, whether statistical or neural, operate on the generated sequence itself. Attention-based attribution maps, by contrast, capture how a proxy Transformer internally processes that sequence, revealing a two-dimensional relational structure over token interactions. This representation can preserve higher-order local regularities that are not directly observable from surface-form statistics alone. We therefore investigate whether these attribution patterns provide a more robust and transferable signal for distinguishing human- and AI-generated text. To test this hypothesis, we introduce AEYEDE, an attribution-based detection framework that operates directly on attention attribution maps extracted from Transformer models (Vaswani et al., 2017). Given an observed text x (human or AI-generated), AEYEDE passes x through a fixed *proxy* model G_θ with white-box access and derives an attention attribution matrix (Sec. 3). We process attribution maps using a multi-scale convolutional encoder with attention pooling to obtain compact embeddings for authorship classification (Figure 1), making the detector less sensitive to purely lexical or stylistic variation.

Beyond predictive performance, this approach also enables an interpretable analysis of the learned attribution structure. For this purpose, we analyze what the CNN attribution encoder captures in attention maps. Clustering 8×8 patches in its last convolutional stage feature space reveals recurring local patterns (*motifs*) whose prevalence differs between human and AI-generated text across datasets and proxy models. This indicates that authorship leaves a localized, repeatable signature in proxy-model attention maps that our detector can exploit.

We evaluate AEYEDE across both encoder-decoder and decoder-only settings, using machine translation benchmarks (WMT14 and the UN Parallel Corpus) and open-ended generation datasets (HC3, RAID, and Beemo). Together, these experiments cover multiple languages, domains, and model families.

Our main contributions are summarized as follows:

- We introduce AEYEDE, an attribution-based framework for AI-generated text detection that uses attention attribution maps from a proxy Transformer as structured input to a lightweight CNN classifier.
- We provide a broad empirical evaluation across encoder-decoder and decoder-only settings, including generator-specific detection, mixed-generator generalization, adversarial perturbations, and cross-dataset transfer. The results show that attention-based attribution maps provide a competitive and complementary detection signal, with particularly strong performance in generator-specific settings and robustness to alternative-spelling attacks.
- We analyze the learned attribution representations and identify recurring local attention patterns, or "motifs," whose relative frequencies differ systematically between human- and AI-generated text. These motifs provide an interpretable and localized signature of authorship.

2 Related Work

AI-Generated Text Detection. Research on detecting machine-generated text has accelerated alongside the rapid progress and deployment of LLMs (Wu et al., 2025). Existing approaches can be categorized into (i) *surface-statistical* and (ii) *likelihood-based* detectors, (iii) *supervised neural classifiers*, (iv) *watermarking and source attribution*, and (v) *LLM-based* meta-detectors. Surface-statistical methods exploit distributional artifacts such as perplexity, burstiness, or n-gram irregularities, often providing lightweight but increasingly fragile signals as generators improve (Gehrmann et al., 2019; Ippolito et al., 2020; Shen et al., 2023; Tassopoulou et al., 2021; Krishna et al., 2022). Complementarily, likelihood-based methods probe the generator's probability landscape: DetectGPT identifies machine-generated text by measuring curvature via perturbations (Mitchell et al., 2023), and related work improves efficiency and robustness through faster perturbation schemes (Bao et al., 2023). These lines of work capture model-specific statistical footprints, but can degrade as LLMs are optimized to match human-like distributions.

Neural Detectors, Robustness, and Generalization. Supervised detectors typically fine-tune Transformer encoders (e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)) on labeled human vs.

machine text, achieving strong in-domain performance but often suffering under domain shift and unseen generators (Uchendu et al., 2021; Wang et al., 2024b). Robustness has become a central focus: training on diverse decoding strategies improves resilience (Ippolito et al., 2020), adversarial training frameworks such as IRON harden detectors against evasion (Li et al., 2025), and Radar explicitly targets robustness via adversarial learning (Hu et al., 2023). Recent methods further aim to improve out-of-distribution behavior and reliability guarantees, e.g., by shaping attention over multiple receptive ranges (Jiao et al., 2025) or bounding false positives with conformal prediction in zero-shot settings (Zhu et al., 2025). Interpretability for detectors is also receiving attention: feature-level analyses using sparse autoencoders help reveal which latent patterns separate machine and human text (Kuznetsov et al., 2025), while downstream applications increasingly require multilingual and domain-specific robustness (Ali et al., 2025) and fine-grained settings such as human-AI co-authorship (Su et al., 2025).

Watermarking and Source Attribution. Watermarking aims to embed detectable signals into generated text, enabling attribution when generation-side cooperation is available (Liu et al., 2024a). Early and widely adopted schemes include token-list or "soft" watermarks that bias sampling (Kirchenbauer et al., 2023), while subsequent work explored alternative embedding mechanisms and detection rules, including entropy- or Bayesian-inspired detectors (Lu et al., 2024; Huang et al., 2025a) and more adaptive watermark designs such as MorphMark (Wang et al., 2025). Recent studies further examine watermark ensembles (Niess & Kern, 2025), watermark-based source attribution (e.g., WASA) (Lu et al., 2025), and approaches that reduce bias and risk (Mao et al., 2025). However, watermarking remains challenged by post-editing and paraphrasing (Liu et al., 2024a), motivating defenses such as paraphrase inversion (Rivera Soto et al., 2025) and robustness through injected "fictitious knowledge" signals (Cui et al., 2025). Adversarial settings also reveal vulnerabilities: DITTO formalizes spoofing attacks against watermarked LLMs via knowledge distillation (Ahn et al., 2025), underscoring the need for evaluation under realistic transformation and attack pipelines.

LLMs as Detectors and Explainable Attribution. Beyond classical detectors, LLMs are increasingly used as meta-detectors and critics of generated content, reflecting a trend toward black-box and instruction-following detection pipelines (Wang et al., 2024b). Recent work expands from binary detection to attribution and explanation, e.g., XDAC provides XAI-driven detection and attribution for Korean news comments (Go et al., 2025a;b), and studies of detectability highlight how author intent and role can affect detection outcomes (Li & Wan, 2025). Together, these directions emphasize that practical detection increasingly requires robustness, reliability, and interpretable evidence—not only raw accuracy.

Benchmarks and Shared Tasks. Progress in AI-text detection is tightly coupled with benchmarks that stress generalization across domains, languages, and attack conditions. Widely used datasets include HC3 (Guo et al., 2023), MGTBench (He et al., 2024), WritingPrompts (Bao et al., 2023), RAID (Dugan et al., 2024), and adversarial extensions such as Adv-HC3 (Peng et al., 2023); additional resources target broader settings such as BUST (Cornelius et al., 2024) and LLMTRACE (Tolstykh et al., 2025). Beyond text-only benchmarks, MultiSocial supports multilingual social-media detection (Macko et al., 2025), Double Entendre introduces a multimodal audio-lyrics setting (Frohmann et al., 2025), and stress-test benchmarks systematically perturb style to probe brittleness (Pedrotti et al., 2025). Shared tasks further standardize evaluation and accelerate methodology: SemEval-2024 Task 8 targets black-box, multilingual, and multidomain detection (Wang et al., 2024a), with system analyses such as TrustAI highlighting practical modeling choices (Urlana et al., 2024). Community efforts such as the GenAIDetect workshop at COLING 2025 (Alam et al., 2025) and domain-focused shared tasks and datasets (e.g., M-DAIGT for news and academic writing (Lamsiyah et al., 2025), and AraGenEval for Arabic settings (Abudalfa et al., 2025)) reflect increasing emphasis on robustness, multilinguality, and real-world constraints. These benchmarks and tasks collectively motivate detectors that generalize across generator families while offering transparent, verifiable evidence for their decisions.

Positioning of Our Work. Our method differs from prior AI-text detectors by using attention-based attribution maps from a proxy Transformer as the main detection signal, rather than relying only on surface statistics, likelihood perturbations, or end-to-end text representations. It does not require watermarking or

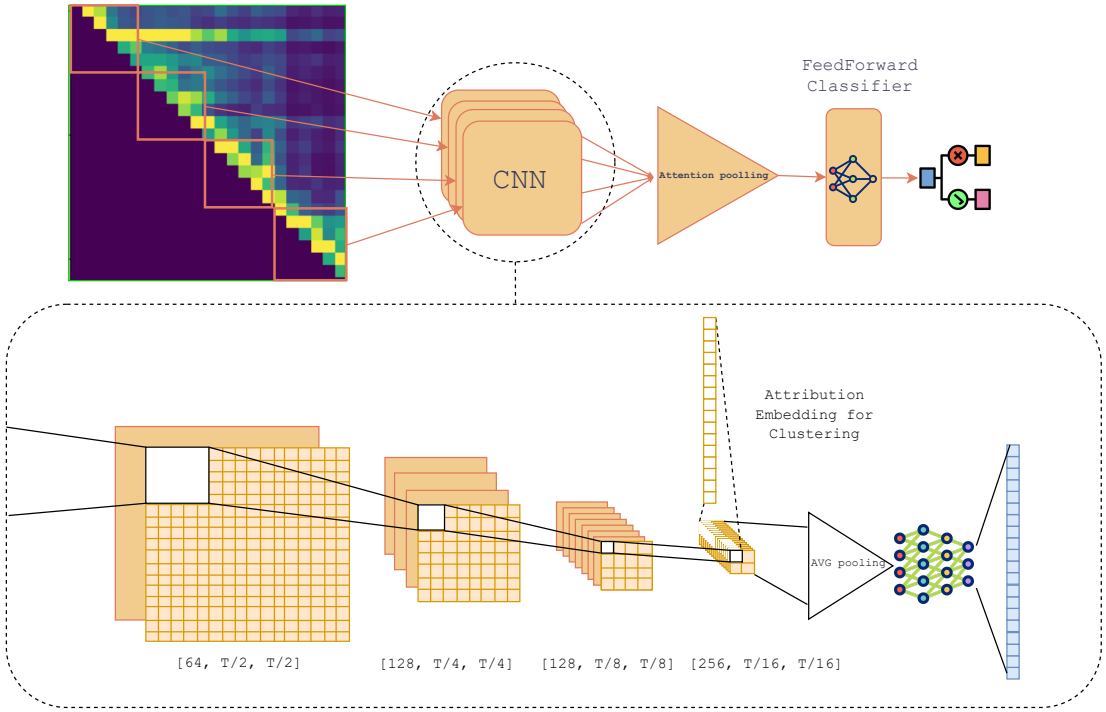


Figure 1: Overview of AEyeDE. Given a text sample and white-box access to a proxy generator model G_θ, we extract an attention-derived attribution matrix A (top left). For decoder-only models, we summarize A by sampling fixed-size square blocks (e.g., 128×128) along the main diagonal, where the strongest local token-token interactions are concentrated (orange boxes) (Xiao et al., 2023; Ivanitskiy et al.; Qi et al., 2025); for encoder-decoder models, we use the full cross-attention map. Each block is encoded by a CNN (bottom), producing one embedding per block. These embeddings are aggregated with learnable attention pooling to form a global attribution representation, which is passed to a lightweight MLP classifier to predict human versus LLM authorship (top right). In our implementation, the final low-resolution CNN feature map is 16×16; its spatial cells correspond to regions of the original attribution block and are later used for motif clustering.

access to the true generator, and it is less dependent on lexical or stylistic cues alone. More specifically, our framework uses attribution maps as structured inputs to a dedicated detection model. The experiments suggest that this signal is useful for this task: it performs especially well in generator-specific settings, remains competitive on standard benchmarks, and shows robustness under alternative-spelling attacks and cross-dataset transfer.

3 AEyeDE Framework

We cast AI-text detection as a binary classification problem. The detector has white-box access to a proxy language model G_θ, which may be either the suspected generator itself or a surrogate model. Given an observed text sample x, we pass x through G_θ and extract an attention-derived attribution map. Within each experiment, the same proxy model is used to compute attributions for all inputs, regardless of whether x is human-written or AI-generated. Our hypothesis is that the internal processing dynamics of G_θ induce systematic differences in these attribution maps for human versus machine-generated text.

For decoder-only models, the attribution map reflects the influence of previous tokens on each current token under causal self-attention. For encoder-decoder models, it reflects the influence of source tokens on each

target token under cross-attention. In both cases, the attribution map serves as the primary input to the downstream detector.

Let $\mathcal{V}$ denote the vocabulary, and let $x = (x_1, \ldots, x_T) \in \mathcal{V}^T$ denote the text whose authorship is to be predicted. Let $y \in \{0, 1\}$ be the corresponding label, where $y = 1$ indicates AI-generated text and $y = 0$ indicates human-written text. When the proxy model is encoder-decoder, the target text x is paired with its associated source sequence to compute cross-attention. The detector is a conditional classifier

$$D_\phi(x; \theta) = f_\phi\big(x, A(x; \theta)\big) \in [0, 1],$$

where $A(x; \theta)$ is the attention-derived attribution map extracted from G_θ, and f_ϕ is a learnable decision function with parameters ϕ.

Assume first that G_θ is an L-layer, H-head Transformer. At layer $\ell \in \{1, \ldots, L\}$ and head $h \in \{1, \ldots, H\}$, let $Q^{(\ell,h)} \in \mathbb{R}^{T_t \times d_k}$ and $K^{(\ell,h)} \in \mathbb{R}^{T_s \times d_k}$ denote the query and key matrices. The attention matrix for that head is

$$\widetilde{A}^{(\ell,h)}(x; \theta) = \mathrm{softmax}\left(\frac{Q^{(\ell,h)}(K^{(\ell,h)})^\top}{\sqrt{d_k}} + C\right) \in \mathbb{R}^{T_t \times T_s},$$

where $C \in \mathbb{R}^{T_t \times T_s}$ is the causal mask in the decoder-only setting. In the encoder-decoder setting, $\widetilde{A}^{(\ell,h)}$ denotes cross-attention with the same target-by-source orientation and no causal mask. We then average across heads and layers to obtain a single attribution map:

$$A(x; \theta) = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \widetilde{A}^{(\ell,h)}(x; \theta) \in \mathbb{R}^{T_t \times T_s}.$$

Each entry

$$a_{t,s} = \big[A(x; \theta)\big]_{t,s}$$

gives the average attention mass assigned from the target position t to source/previous position s under the proxy model. In the decoder-only setting, $T_s = T_t = T$; in the encoder-decoder setting, T_s and T_t may differ. We use the `Inseq` library to extract these attention-derived attributions (Sarti et al., 2023).

The detector architecture is shown in Fig. 1. For decoder-only models, we summarize A by extracting square blocks of size w_a along a diagonal traversal; in all experiments, we set $w_a = 128$. In encoder-decoder experiments, the inputs are limited to at most 128 tokens, so we use a single block that covers the full cross-attention map.[1] This design is motivated by the empirical observation that, in decoder-only models, the most informative attribution mass is concentrated near the main diagonal (Xiao et al., 2023; Ivanitskiy et al.; Qi et al., 2025).

Let (p_k, q_k) denote the top-left corner of the k-th block. We define

$$A_k = A[p_k : p_k + w_a,\ q_k : q_k + w_a] \in \mathbb{R}^{w_a \times w_a}. \tag{1}$$

Let $m_t \in \{0, 1\}^{T_t}$ and $m_s \in \{0, 1\}^{T_s}$ denote the target- and source-side padding masks, respectively, and define the validity mask

$$P = m_t m_s^\top \in \{0, 1\}^{T_t \times T_s}, \tag{2}$$

which marks non-padding attribution entries. A block is considered valid if it contains at least one non-padding entry:

$$u_k = \mathbb{I}\left(\sum_{i,j} P[p_k + i,\ q_k + j] > 0\right) \in \{0, 1\}. \tag{3}$$

Each valid block is encoded by a CNN-based attribution encoder

$$b_k = E_{\mathrm{attr}}(A_k) \in \mathbb{R}^{d_{\mathrm{attr}}}. \tag{4}$$

[1]For example, on HC3 (see Section 4.1), the 128×128 diagonal band exhibits higher entropy than the off-diagonal region (8.89 vs. 7.78), together with a higher mean attribution value (0.007 vs. 0.001) and a larger standard deviation (0.01 vs. 0.002). Processing four 128×128 patches is also considerably more efficient than processing a full 1024×1024 matrix.

The attribution encoder E_{attr} treats A_k as a single-channel image. It applies a sequence of convolutional blocks with channel progression

$$1 \to 32 \to 64 \to 128 \to 128 \to 256,$$

interleaved with 2×2 max-pooling, followed by global average pooling and a two-layer MLP. Writing $g(\cdot)$ for the convolutional pipeline plus global average pooling, we obtain

$$r_k = g(A_k) \in \mathbb{R}^{256}, \tag{5}$$

and then

$$b_k = W_2\, \rho\big(\mathrm{BN}(W_1 r_k + b_1)\big) + b_2, \tag{6}$$

where $\rho(\cdot)$ denotes ReLU. In our implementation, the final convolutional feature map is 16×16; its spatial cells can be mapped back to local regions of the original attribution block for motif clustering (see Sec. 5).

Optionally, we augment the attribution representation with a text representation.[2] To handle long sequences, we pad the input to a multiple of a fixed window size w_t and partition it into

$$N = \left\lceil \frac{T}{w_t} \right\rceil$$

contiguous chunks:

$$x^{(i)} = x_{(i-1)w_t+1:iw_t}, \qquad i = 1, \ldots, N. \tag{7}$$

A text encoder E_{text} maps each chunk to a vector

$$c_i = E_{\text{text}}\big(x^{(i)}\big) \in \mathbb{R}^{d_{\text{text}}}. \tag{8}$$

We ignore fully padded chunks via a validity indicator $v_i \in \{0, 1\}$.

We aggregate sets of per-block vectors using learnable attention pooling. Given vectors $\{z_i\}_{i=1}^n$ and a corresponding validity mask $\{m_i\}_{i=1}^n$, we define

$$s_i = \omega^\top \tanh(W_p z_i), \tag{9}$$

$$\alpha_i = \frac{m_i \exp(s_i)}{\sum_j m_j \exp(s_j)}, \tag{10}$$

$$\mathrm{Pool}(\{z_i\}, \{m_i\}) = \sum_i \alpha_i z_i, \tag{11}$$

where ω and W_p are learnable parameters. Applying this operator to the text and attribution branches gives

$$h_{\text{text}} = \mathrm{Pool}(\{c_i\}, \{v_i\}), \qquad h_{\text{attr}} = \mathrm{Pool}(\{b_k\}, \{u_k\}). \tag{12}$$

The final representation is

$$h = \begin{cases} [\, h_{\text{text}};\, h_{\text{attr}} \,], & \text{if the text branch is used,} \\ h_{\text{attr}}, & \text{otherwise.} \end{cases} \tag{13}$$

A two-layer MLP then produces a scalar logit ℓ and the corresponding probability of AI authorship:

$$\ell = w^\top \rho(W_h h + b_h) + b, \tag{14}$$

$$p(y = 1 \mid x, A) = \sigma(\ell). \tag{15}$$

We train the detector using binary cross-entropy on labeled examples.

In this work, prompts are discarded in the decoder-only setting. Thus, the detector conditions only on the observed continuation $x_{1:T}$ and the attribution map derived from it, rather than on the full prompt-response pair.

[2] We evaluate this text-augmented variant for encoder-decoder models. In practice, attention-only representations are consistently competitive with the text-augmented variant, so we emphasize the attention-based results in the main paper.

4 Experimental Results

4.1 Datasets

In this study, we evaluate both encoder–decoder and decoder-only language models. For the encoder–decoder setting, we conduct experiments on three translation language pairs: French–English (fr-en) and German–English (de-en) using WMT14 (Bojar et al., 2014), and Arabic–English (ar-en) using the UN Parallel Corpus (Ziemski et al., 2016). For each language pair, we construct a dataset of 200k source–target examples consisting of gold (human) reference translations and corresponding model-generated translations produced by the Marian-MT model (Tiedemann & Thottingal, 2020). We selected samples with source and target pairs of at most 128 tokens.

For the decoder-only setting, we use the HC3 dataset (Guo et al., 2023), RAID (Dugan et al., 2024), and Beemo (Artemova et al., 2025). HC3 provides paired human and ChatGPT (OpenAI, 2023) responses. We remove examples exceeding 1,024 tokens and retain approximately 24k samples per class. RAID contains human-written text and model-generated text spanning multiple domains. From RAID, we use outputs from Cohere, Llama 2 70B (chat), GPT-2 XL, and Mistral-7B, together with the corresponding human responses to the same prompts. Because RAID is imbalanced (with fewer human than model-generated examples), we downsample each model's generated subset to 26,700 examples (the minimum across selected models) and use all available human examples (12,900). We use Beemo as an external test dataset for cross-dataset out-of-distribution experiments. With our preprocessing on Beemo, we obtain 2,178 human-written samples and an equal number of machine-generated samples, which are sourced from multiple LLMs. In all experiments, we use a class-balanced entropy loss to mitigate the imbalance between human- and AI-generated samples.

We split each dataset into 90% training, 5% validation, and 5% test. Throughout our experiments with decoder-only models, we discard prompts, as in practice, it is more likely that a text sample is observed without access to the prompt that generated it. We use Llama 3.1 8B (Grattafiori et al., 2024), Cohere (c4ai-command-r7b-12-2024)(Cohere et al., 2025), GPT-neo (Black et al., 2021), and Mistral (Ministral-3-8B-Instruct-2512) (Jiang et al., 2023) to obtain the attributions from approximate models of the same family as those that generated the RAID and HC3 datasets.

4.2 Evaluation Metrics

For evaluation, we report Accuracy, Precision, Recall, F1, Area Under Curve, and True/False-positive Rate at a fixed low false-positive operating point, namely TPR@FPR = 0.01. Here $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$; thus FPR = 0.01 corresponds to falsely labeling 1% of human-written samples as AI-generated. This is the critical "high-precision" regime. In academic or forensic settings, false positives (accusing a human of using AI) are unacceptable. A detector must have a high TPR at a very low FPR to be deployable (Ayoobi et al., 2025). For all threshold-dependent metrics, thresholds were selected on the validation split and then fixed for test evaluation.

4.3 Baseline Models

For the encoder-decoder setting, our baseline is a custom three-layer Transformer-based classifier trained on paired source-target text. For the decoder-only setting, we report results from Fast-DetectGPT (Curvature) (Bao et al., 2023), Binoculars (Hans et al., 2024), Rank and LogRank (Su et al., 2023), GLTR (Gehrmann et al., 2019), and a RoBERTa-based detector released by SuperAnnotate.[3]

Curvature compares the conditional log-probability of each observed token with that of plausible alternative tokens in the same context, and aggregates these comparisons into a sample-level statistic called conditional probability curvature. Machine-generated text tends to show a higher curvature value because its token choices are more often locally optimal under the model. Binoculars computes the ratio between a model's own perplexity and its cross-perplexity with another model. The intuition is that machine-generated text is not only low-perplexity for an LLM, but also shows unusually high agreement across related models[4]. GLTR

[3] `https://huggingface.co/SuperAnnotate/ai-detector`

[4] We use LLama-3.1-8B and Llama-3.1-8 B-Instruct the pair models.

detects machine-generated text by scoring each token under a reference language model using its probability, rank, and predictive entropy. Rank is a token-level measure that extracts the rank of each token under a language model and averages it across the sequence: lower Rank values indicate that the sample is more likely to be machine-generated. LogRank computes the logarithm of this rank-based score.

4.4 Results

We evaluate the proposed framework across five experimental settings, followed by an analysis of attribution-region motifs, each of which tries to answer a research question of the use of attention attribution for this task. First, we apply AEyeDE to encoder-decoder attention maps and examine the effect of combining attribution maps with the text signal. Second, we assess in-domain detection performance, where the proxy model matches the generator of the evaluated text. Third, we consider a unified setting in which samples from multiple LLMs are mixed during training, and generalization includes a held-out generator. Fourth, we evaluate robustness under adversarial attacks. Fifth, we examine cross-dataset generalization on an external benchmark.

Do attribution maps provide useful detection signals in encoder-decoder models? In encoder–decoder architectures, cross-attention models the dependencies between source and target tokens by enabling the decoder to focus selectively on the most relevant encoder states during generation. In our task, we use the entire 128×128 attention matrix as the signal to be processed (as we selected samples of at most 128 tokens). Table 1 reports results for AI-translation detection across three Marian MT-language pairs. Using attribution maps alone (CNN), our method consistently outperforms the text-only baseline for all metrics, with gains of +3.6 F1 for ar-en (74.6 vs. 71.0), +6.7 for de-en (81.5 vs. 74.8), and +8.3 for fr-en (85.1 vs. 76.8). Adding target-text features (CNN+text) yields a further, but smaller, improvement over CNN in every case (+2.0, +1.7, and +1.4 F1, respectively), suggesting that most discriminative signal is already captured by the attribution structure, while text provides complementary information.

Table 1: Performance on Marian-MT generated translation and attribution maps. CNN and CNN+text configurations are based on AEyeDE.

	ar-en					**de-en**					**fr-en**				
Config	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
AEyeDE(CNN)	68.9	63.0	91.4	74.6	77.7	79.0	72.6	93.0	81.5	87.3	83.6	78.0	93.6	85.1	91.0
AEyeDE(CNN+text)	**71.9**	**65.7**	**91.9**	**76.6**	**81.5**	**81.1**	**74.9**	**93.6**	**83.2**	**89.2**	**85.4**	**80.3**	**93.9**	**86.5**	**92.7**
text (baseline)	61.9	57.3	93.4	71.0	71.5	69.5	63.7	90.5	74.8	77.5	72.4	66.2	91.4	76.8	81.4

How effective is AEyeDE when the proxy model matches the generator family? Using the RAID dataset, we train a separate detector for each generator family using only samples generated by that family, along with the corresponding matched human texts (e.g., we train Llama samples with the attributions extracted from the Llama model). We refer to this configuration as the *individual* setting. For the model-specific baselines (Curvature, GLTR, Rank, and LogRank), the underlying scores are also extracted from the corresponding generator model. It should be noted that both LLMs used for binoculars are from Llama family.

In the *individual* setting (Table 2), AEyeDE achieves near-ceiling performance across all four generator families, with F1 ranging from 96.34 to 99.25 and AUC from 98.81 to 99.86, and it is the top-performing method on every generator. The gains are especially large for GPT-neo and Mistral. For GPT-neo, the strongest baseline reaches 83.70 F1 (RoBERTa), whereas AEyeDE attains 97.96. For Mistral, the strongest baseline reaches 82.71 F1 (RoBERTa), compared with 96.34 for AEyeDE. On Cohere, the strongest baselines are more competitive, with RoBERTa reaching 88.17 F1 and Curvature 87.81, but AEyeDE still improves substantially to 97.95. The baseline behavior is more heterogeneous for Llama. Several methods perform strongly in this case, including Binoculars (98.58 F1, 99.66 AUC), RoBERTa (96.98 F1), and GLTR (90.70 F1). Even under this more favorable setting for the baselines, AEyeDE remains best overall, achieving 99.25 F1, 99.86 AUC, and 99.33 on the low-FPR operating-point metric.

Table 2: Results on *individual* arrangement where each human and AI-generated attention attributions were extracted from the same model specified in the dataset to train the AEYEDE model.

	Cohere						GPT-neo					
	Acc	Prec	Rec	F1	AUC	TPR-FPR	Acc	Prec	Rec	F1	AUC	TPR-FPR
AEYEDE	**97.23**	**97.48**	98.43	**97.95**	**99.51**	**95.22**	**97.22**	97.20	98.73	**97.96**	**99.71**	**95.89**
Binoculars	67.94	67.91	99.48	80.72	78.10	34.83	67.61	67.63	**99.78**	80.62	25.84	0.45
Curvature	83.27	86.29	89.39	87.81	90.86	47.01	78.91	**98.22**	70.03	81.76	79.70	65.47
GLTR	67.39	67.42	**99.93**	80.52	76.68	10.99	67.96	67.92	99.55	80.75	71.54	44.99
LogRank	67.39	67.42	**99.93**	80.52	75.49	10.24	69.58	69.31	98.58	81.39	73.95	36.70
Rank	67.39	67.42	**99.93**	80.52	53.66	0.00	73.11	77.14	85.50	81.11	79.05	21.52
RoBERTa	83.27	84.26	92.45	88.17	90.42	23.92	75.13	75.04	94.62	83.70	79.58	20.03
	Llama						Mistral					
	Acc	Prec	Rec	F1	AUC	TPR-FPR	Acc	Prec	Rec	F1	AUC	TPR-FPR
AEYEDE	**98.99**	**99.18**	99.33	**99.25**	**99.86**	**99.33**	**95.09**	**97.40**	95.29	**96.34**	**98.81**	**88.94**
Binoculars	98.08	98.80	98.36	98.58	99.66	97.91	67.48	67.63	99.63	80.57	55.49	11.81
Curvature	67.37	67.44	99.85	80.51	78.67	39.69	67.68	67.68	**100.00**	80.72	70.32	49.10
GLTR	87.39	90.30	91.11	90.70	93.30	41.18	67.68	67.68	**100.00**	80.72	69.29	36.77
LogRank	85.93	90.08	88.94	89.51	92.16	38.04	67.93	68.00	99.40	80.75	69.01	28.03
Rank	67.57	67.54	**100.00**	80.63	51.41	0.00	67.83	67.80	99.93	80.79	59.71	0.00
RoBERTa	95.97	97.94	96.04	96.98	98.35	83.86	72.89	72.76	95.81	82.71	79.17	11.36

Table 3: Results on the *unified* arrangement where the samples corresponding to the Mistral model were excluded from the training of AEYEDE.

	Cohere						GPT-neo					
	Acc	Prec	Rec	F1	AUC	TPR-FPR	Acc	Prec	Rec	F1	AUC	TPR-FPR
AEYEDE	**72.38**	90.98	65.55	76.19	**80.67**	**58.67**	**72.25**	**96.57**	61.06	74.82	**78.86**	**58.97**
Binoculars	67.29	67.57	99.03	80.33	74.21	49.33	67.46	67.51	99.85	80.55	74.43	43.57
Curvature	57.41	79.03	50.15	61.36	64.00	20.18	70.13	93.37	60.01	73.07	77.82	52.84
GLTR	67.44	67.44	**100.00**	**80.55**	76.32	26.53	67.51	68.11	97.53	80.21	73.42	15.10
LogRank	67.44	67.44	**100.00**	**80.55**	75.14	24.51	68.06	69.44	94.10	79.91	73.20	16.52
Rank	67.44	67.44	**100.00**	**80.55**	47.76	0.52	67.51	67.51	**100.00**	**80.60**	64.02	4.11
RoBERTa	69.66	80.87	72.05	76.21	77.34	35.50	69.48	75.83	80.42	78.06	77.17	38.27
	Llama						Mistral					
	Acc	Prec	Rec	F1	AUC	TPR-FPR	Acc	Prec	Rec	F1	AUC	TPR-FPR
AEYEDE	**71.96**	**92.59**	63.53	75.35	**79.29**	**56.80**	**71.83**	**95.14**	61.51	74.72	**78.86**	**58.59**
Binoculars	67.52	67.61	99.55	80.53	75.99	50.45	67.48	67.70	99.33	80.52	72.59	46.49
Curvature	63.04	82.42	57.47	67.72	69.42	27.50	59.28	81.84	51.20	62.99	65.14	28.48
GLTR	67.47	67.47	**100.00**	**80.58**	74.20	22.12	67.63	67.66	99.93	80.69	75.19	24.07
LogRank	67.47	67.47	**100.00**	**80.58**	72.75	18.91	67.68	67.68	**100.00**	**80.72**	74.63	20.55
Rank	67.47	67.47	**100.00**	**80.58**	45.44	0.52	67.68	67.68	**100.00**	**80.72**	54.88	0.22
RoBERTa	70.20	77.48	78.70	78.09	77.92	39.76	71.12	79.89	76.61	78.21	78.49	38.86

Table 4: Results on HC3 using GPT-neo to extract attention attributions.

	GPT-neo					
	Acc	Prec	Rec	F1	AUC	TPR-FPR
AEYEDE	96.85	95.68	98.27	96.96	99.51	93.50
Curvature	**98.40**	**99.83**	97.04	**98.42**	99.22	97.45
GLTR	98.25	98.89	97.60	98.24	99.17	97.60
LogRank	97.95	99.69	96.20	97.91	99.03	97.50
Rank	83.40	81.63	86.20	83.85	91.22	36.60
RoBERTa	97.73	96.70	**98.93**	97.80	**99.57**	**97.78**

Furthermore, a broader pattern is that likelihood- and rank-based baselines, especially on Cohere, GPT-neo, and Mistral, yield recall near 100% but precision close to the class prior ($\sim$ 67%). In contrast, AEyeDE maintains both high precision and high recall across all generators, which suggests that the attention-based attribution signals are highly discriminative when the proxy model matches the generator family.

Likewise, we train and test AEyeDE on the HC3 dataset. Table 4 reports results on the HC3 dataset using attribution maps extracted from GPT-neo as the proxy model, as it is generated only by the GPT family. Overall, HC3 appears substantially less challenging than RAID, as nearly all methods achieve very high performance. AEyeDE remains competitive, achieving 96.85 accuracy, 95.68 precision, 98.27 recall, 96.96 F1, and 99.51 AUC. Although Curvature obtains the best F1 and RoBERTa achieves the highest AUC, AEyeDE is notable because it reaches near-ceiling performance using a fundamentally different signal. Moreover, the comparison with RoBERTa should be interpreted cautiously, since HC3 was part of that model's training data. We employ the trained models on the RAID dataset along with the model trained on HC3 data, to analyze informative motif patterns for human vs. AI-generated text (See 5).

Do attribution-based representations transfer to unseen generators? In a different setup, we evaluate how well our proposed model can generalize to an unseen generator at test time. We train on a *unified* mixture of all RAID generators except Mistral, but the test set includes Mistral's generation. To control for training set size, we subsample each included generator's generated text to match the per-model training budget used in the *individual* setting. This splitting strategy assesses whether representations learned from a subset of generator families transfer to an unseen model at test time.

In the *unified* setting (Table 3), performance is lower than in the generator-matched *individual* setting, reflecting the increased heterogeneity of pooled training and the greater difficulty of cross-generator transfer. Even so, AEyeDE achieves the best accuracy, AUC, and TPR@FPR=0.01 for every generator family. Its AUC ranges from 78.86 to 80.67, and its TPR@FPR=0.01 ranges from 56.80 to 58.97, consistently exceeding all baselines. At the same time, AEyeDE adopts a markedly more conservative operating point than most competing methods. Across all generators, it maintains very high precision (90.98-96.57) but substantially lower recall (61.06-65.55), which leads to F1 scores in the 74.72-76.19 range. In contrast, Binoculars, GLTR, LogRank, and Rank operate in an almost-all-positive regime, with recall near 100% but precision close to the class prior ($\sim$ 67%). Their F1 values therefore appear superficially strong ($\sim$ 80.2-80.7), but this comes at the cost of many false positives. RoBERTa is the strongest balanced baseline, with F1 between 76.21 and 78.21 and AUC between 77.17 and 78.49, yet it still trails AEyeDE on the threshold-independent metrics and on the low-FPR operating point.

These results suggest that attribution-based representations do transfer across generator families, but they do so conservatively: when trained on mixed generators, AEyeDE becomes more selective, flagging fewer human texts incorrectly while missing a larger fraction of AI-generated samples at the default threshold. Importantly, its consistent advantage in AUC and TPR@FPR=0.01 indicates that it provides the best overall separation between human and machine text under cross-generator transfer.

How robust is AEyeDE to adversarial perturbations? AI-generated text detectors are known to be brittle under adversarial perturbations (Krishna et al., 2023). We evaluate robustness to the paraphrasing and alternative spelling attacks provided in the RAID dataset for both the individual and unified settings.[5]

In the individual setting (Table 5), AEyeDE achieves the highest F1 (81.50) and AUC (77.64), with a substantial precision advantage over all baselines (74.78 vs. 68.87 for the next-best method, RoBERTa)[6]. Binoculars, GLTR, LogRank, and Rank exhibit near-perfect recall ($\geq$ 99.48%) but precision close to the class prior (~67%), indicating a near-all-positive prediction behaviour. Curvature attains the highest accuracy (76.13), but its recall drops sharply (54.13), yielding by far the lowest F1 (55.69). In contrast, AEyeDE maintains the best precision-recall tradeoff (74.78 / 89.54), suggesting that attribution-based structure appears more robust to paraphrastic surface perturbations in the individual setting.

[5]Due to computational constraints, we report these adversarial experiments for GPT-neo only.

[6]The paraphrases are generated by a T5 model, so the adversarial samples remain machine-generated, although they are out-of-distribution relative to the training data. Importantly, adversarial samples are used only at test time, meaning that AEyeDE is not trained on adversarial examples.

In the unified adversarial setting (Table 6), the distinction between methods becomes much smaller in terms of thresholded metrics: accuracy remains near the majority-class baseline (~67%) and F1 values cluster tightly (78.92-80.45). AEyeDE still attains the highest F1 (80.45), but only by a narrow margin, and its precision/recall profile (67.50 / 99.55) indicates that it too shifts toward the near-all-positive trend. Moreover, its AUC drops to 62.21, below RoBERTa (72.87) and Curvature (70.38). Overall, these results show that paraphrasing remains a difficult adversarial condition for all detectors. AEyeDE retains a clear advantage in the individual setting, but that advantage disappears in the unified setting.

We also evaluate robustness to alternative-spelling perturbations, motivated by the fact that spelling errors are relatively uncommon in clean LLM outputs. In the RAID dataset, this attack is constructed by randomly modifying characters in the original samples. In the individual setting (Table 7), AEyeDE remains highly effective under this attack, achieving the best F1 (98.47), AUC (99.75), and TPR@FPR=0.01 (96.26). It outperforms all baselines by wide margins, with the next-best F1 being 84.06 for RoBERTa. Several baselines, especially Binoculars and GLTR, operate close to an all-positive score, with near-perfect recall but substantially weaker precision and poor low-FPR detection. These results suggest that character-level spelling perturbations affect surface-level statistics more than the attribution patterns used by AEyeDE.

In the unified setting (Table 8), performance decreases across methods, but AEyeDE still achieves the best accuracy (70.34), F1 (81.33), AUC (83.48), and TPR@FPR=0.01 (59.10). Its low-FPR true-positive rate exceeds that of the next-best method, Curvature (48.87), by more than 10 points. Curvature attains the highest precision (93.88), but its recall drops to 56.47, yielding the lowest F1 among the reported methods (70.52). Overall, these results suggest that attribution-based detection particularly works well against this type of character-level perturbation. One possible explanation is that subword tokenization does not fully obscure the token-level relationships captured by the attribution maps.

Table 5: Results under the RAID paraphrasing adversarial in the *individual* setting with GPT-neo.

Method	**GPT-neo**					
	Acc	Prec	Rec	F1	AUC	TPR
AEyeDE	72.68	**74.78**	89.54	**81.50**	**77.64**	23.17
Binoculars	67.20	67.20	**100.00**	80.38	15.15	0.00
Curvature	**76.13**	57.34	54.13	55.69	72.38	26.67
GLTR	67.05	67.17	99.70	80.26	65.95	26.98
LogRank	67.10	67.19	99.78	80.30	66.32	**28.48**
Rank	66.90	67.12	99.48	80.16	68.99	13.30
RoBERTa	68.96	68.87	98.21	80.96	68.73	5.08

Table 6: Results under the RAID paraphrasing adversarial in the *unified* setting with GPT-neo.

Method	**GPT-neo**					
	Acc	Prec	Rec	F1	AUC	TPR
AEyeDE	**67.52**	67.50	99.55	**80.45**	62.21	4.80
Binoculars	67.12	67.12	**100.00**	80.33	66.26	18.98
Curvature	67.02	67.12	99.70	80.23	70.38	**19.88**
GLTR	67.17	67.15	**100.00**	80.35	66.77	6.98
LogRank	67.17	67.15	**100.00**	80.35	65.50	10.65
Rank	67.12	67.12	**100.00**	80.33	57.02	2.55
RoBERTa	66.21	**67.89**	94.22	78.92	**72.87**	16.58

Table 7: Results under the RAID alternative-spelling adversarial in the *individual* setting with GPT-neo.

Method	**GPT-neo**					
	Acc	Prec	Rec	F1	AUC	TPR
AEyeDE	**97.93**	**98.07**	98.88	**98.47**	**99.75**	**96.26**
Binoculars	67.44	67.44	**100.00**	80.55	33.56	2.02
Curvature	77.37	96.25	69.13	80.47	77.02	64.65
GLTR	67.39	67.42	99.93	80.52	69.39	42.38
LogRank	68.65	68.92	97.46	80.74	72.28	32.36
Rank	74.55	77.93	86.85	82.15	78.11	15.47
RoBERTa	76.21	76.66	93.05	84.06	80.56	6.28

Table 8: Results under the RAID alternative-spelling adversarial in the *unified* setting with GPT-neo.

Method	**GPT-neo**					
	Acc	Prec	Rec	F1	AUC	TPR
AEyeDE	**70.34**	70.58	95.94	**81.33**	**83.48**	**59.10**
Binoculars	67.30	67.32	99.92	80.44	70.60	37.59
Curvature	68.22	**93.88**	56.47	70.52	75.00	48.87
GLTR	64.78	78.98	64.96	71.29	69.55	16.17
LogRank	65.18	80.00	64.36	71.33	69.77	14.51
Rank	67.31	67.31	**100.00**	80.46	67.18	4.21
RoBERTa	69.74	78.33	76.09	77.19	76.00	8.05

Does AEyeDE generalize to a fully external dataset? To complement the adversarial evaluations and assess generalization on a fully external test set, we evaluate the methods on Beemo (Table 9), which is unseen during training, using Mistral-7B-Instruct as the proxy model [7]. AEyeDE achieves the highest accuracy (66.36), precision (62.16), F1 (71.32), and AUC (74.43), outperforming all other methods on these metrics. Several baselines exhibit thresholded behavior close to an all-positive classifier, with accuracy near

[7] In our experiments, Mistral scored best among the other generators for this task.

50-51%, precision near the class prior (~50%), and recall above 96%, indicating poor calibration at the chosen operating point. RoBERTa performs somewhat better (56.03 accuracy, 68.66 F1) but still trails AEyeDE by 2.66 F1 points. Notably, GLTR achieves a nearly identical AUC (74.38) and the highest TPR@FPR=0.01 (15.46, compared with 14.55 for AEyeDE), although its accuracy and precision remain close to chance at the selected threshold. Overall, these results indicate that the attribution-based signal transfers well across datasets and provides the strongest overall cross-dataset performance, even though some baselines remain competitive on ranking-based metrics such as AUC or low-FPR TPR. It is also noteworthy that these results are obtained with a relatively small training set, suggesting that the proposed framework can extract useful discriminative structure even under limited-data conditions.

Table 9: Cross-dataset evaluation on Beemo using Mistral as the proxy model.

Method	**Mistral**					
	Acc	Prec	Rec	F1	AUC	TPR
AEyeDE	**66.36**	**62.16**	83.64	**71.32**	**74.43**	14.55
Binoculars	51.36	50.71	96.36	66.45	63.16	6.36
Curvature	49.44	49.74	98.73	66.16	50.47	9.76
GLTR	51.40	50.74	**98.88**	67.07	74.38	**15.46**
LogRank	51.27	50.68	98.47	66.92	74.25	13.01
Rank	50.92	50.50	97.71	66.59	64.16	3.05
RoBERTa	56.03	53.37	96.24	68.66	69.17	1.02

5 Analysis of Patch-Level Motifs in Attribution Maps

Building on the attribution encoder E_{attr} (Sec. 3), we investigated whether the detector exploits *localized* and *repeatable* visual motifs in attribution maps that are characteristic of human-written versus AI-generated text. Recall that after the final convolutional stage, each block $A_k \in \mathbb{R}^{128\times128}$ is mapped to a feature map of spatial size 16×16 with 256 channels. We denote this last feature map by

$$F_k \in \mathbb{R}^{256\times16\times16}.$$

Because $128/16 = 8$, each feature map cell (u, v) corresponds to an 8×8 patch of the original block A_k (shown in Fig. 1). Let $P_{k,u,v} \in \mathbb{R}^{8\times8}$ be this patch, and define its representation, obtained after the last CNN convolutional stage, as:

$$z_{k,u,v} \;=\; F_k[:, u, v] \in \mathbb{R}^{256}.$$

We denote $\{z_{k,u,v}\}$ as an embedding space of patches produced by the detector model.

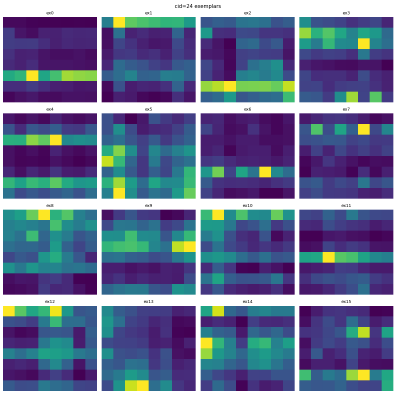
Patch selection and clustering. To avoid padding-only and non-informative constant patches, we only keep patches that (i) correspond to non-padding entries under the mask $M = m_x m_y^\top$, and (ii) pass a minimal informativeness threshold based on mean and standard deviation inside the patch. We keep patches with $|\mu_{k,u,v}| > \tau_\mu(0.01)$ and $\sigma_{k,u,v} > \tau_\sigma(0.001)$. Then, we cluster their retained embeddings $\{z_{k,u,v}\}$ using HDBSCAN, producing assignments of each patch to the corresponding cluster

$$c_{k,u,v} \in \{1, \ldots, C\} \cup \{-1\}. \tag{16}$$

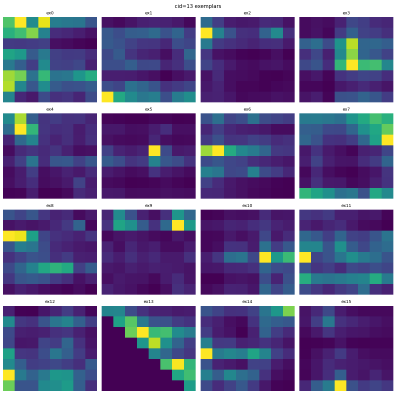
where -1 is unclustered noise. Each cluster can be interpreted as a *motif family*: a set of patches that the trained encoder maps to nearby representations.

Sample-normalized motif rates. A direct comparison of raw motif counts is complicated by two sources: (i) datasets can be class-imbalanced (e.g., RAID), and (ii) the number of valid patches retained after preprocessing varies greatly across samples due to length and informativity selection. To obtain a metric that is comparable across samples, for each sample s we denote by $\mathcal{P}(s)$ the set of all retained patches extracted from that sample (across all diagonal blocks), and define the per-sample *motif rate* for cluster c as

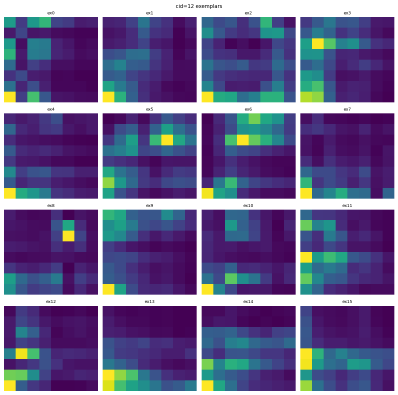
$$r_s(c) \;=\; \frac{1}{|\mathcal{I}_s|} \sum_{(k,u,v)\in\mathcal{I}_s} \mathbb{I}[c_{k,u,v} = c]\,, \tag{17}$$



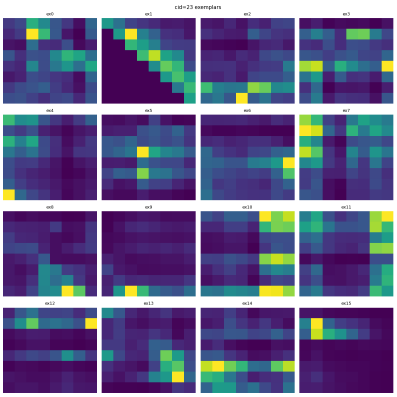
(a) HC3 (individual), proxy: GPT-neo. Top cluster by $\Delta\bar{r}$ (machine-skewed).



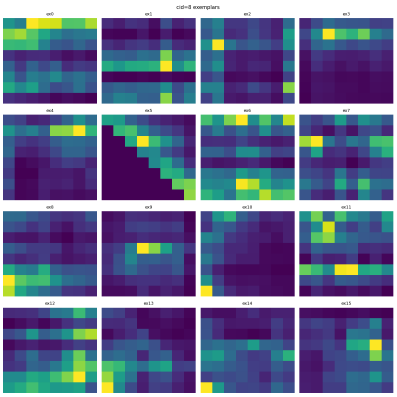
(b) RAID (individual), proxy: Cohere. Top cluster by $\Delta\bar{r}$ (human-skewed).



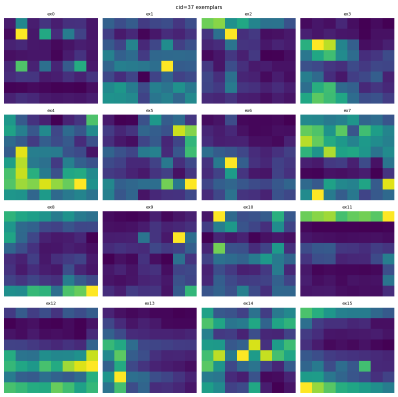
(c) RAID (individual), proxy: Llama. Top cluster by $\Delta\bar{r}$ (human-skewed).



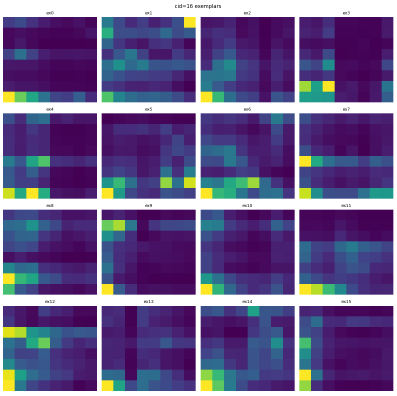
(d) RAID (individual), proxy: Mistral. Top cluster by $\Delta\bar{r}$ (human-skewed).



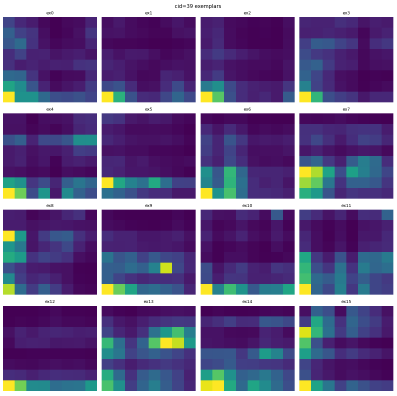
(e) RAID (individual), proxy: Cohere. Top cluster by $\Delta\bar{r}$ (human-skewed).



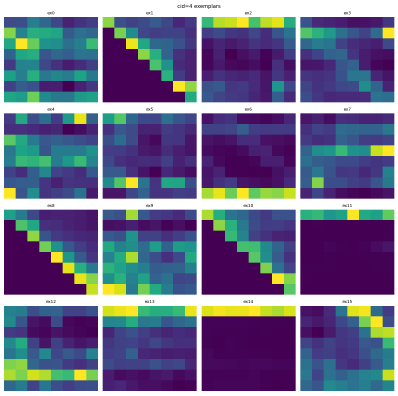
(f) RAID (unified), proxy: GPT-neo. Top cluster by $\Delta\bar{r}$ (human-skewed).



(g) RAID (unified), proxy: Llama. Top cluster by $\Delta\bar{r}$ (human-skewed).



(h) RAID (unified), proxy: Mistral. Top cluster by $\Delta\bar{r}$ (human-skewed).



(i) RAID (individual), proxy: GPT-neo. Top cluster by $\Delta\bar{r}$ (human-skewed).

Figure 2: Examples of the top motif cluster (by absolute mean prevalence gap $\Delta\bar{r}$ between gold-machine and gold-human) for each dataset/proxy-model configuration. Each panel shows representative 8×8 patches (z-normalized) from the corresponding cluster.

Table 10: Top-1 motif cluster by absolute class-rate gap $|\Delta r(c)|$ for each dataset/setting. Rates are mean per-sample motif shares (%), and Δr is reported in percentage points (machine minus human). Confidence intervals are 95% bootstrap CIs over samples; p is from a label-permutation test.

Dataset	**Setting**	**Proxy-model**	**cid**	**Human $\bar{r}_g$**	**Machine $\bar{r}_g$**	Δr	**95% CI**	p
HC3	individual	GPT-neo 2.7B	24	9.7	20.7	+11.0	[9.3, 12.7]	<0.001
RAID	unified	Cohere Cmd-R 7B	13	80.0	69.5	-10.5	[-12.9, -8.2]	<0.001
RAID	individual	Mistral 7B	23	63.5	54.8	-8.7	[-11.1, -6.2]	<0.001
RAID	individual	GPT-neo 2.7B	4	87.6	81.1	-6.5	[-9.0, -4.1]	<0.001
RAID	individual	Llama 3.1 8B	12	7.5	3.1	-4.4	[-4.8, -3.9]	<0.001
RAID	unified	GPT-neo 2.7B	37	11.6	8.5	-3.0	[-4.1, -2.0]	<0.001
RAID	unified	Mistral 7B	39	7.3	5.1	-2.2	[-2.9, -1.6]	<0.001
RAID	individual	Cohere Cmd-R 7B	8	82.1	79.9	-2.1	[-3.7, -0.6]	0.031
RAID	unified	Llama 3.1 8B	16	9.5	7.6	-2.0	[-2.4, -1.5]	<0.001

Here $\mathcal{I}_s$ is the set of retained patch indices (k, u, v) from sample s, and let $c_{k,u,v}$ be the patch-cluster assignment. That is, $r_s(c)$ is the fraction of all retained patches in sample s that belong to cluster c.

For a group of samples $\mathcal{S}_g$ (defined below), we summarize prevalence by the mean rate

$$\bar{r}_g(c) \;=\; \frac{1}{|\mathcal{S}_g|} \sum\nolimits_{s \in \mathcal{S}_g} r_s(c). \tag{18}$$

In our analysis, we report $\bar{r}_g(c)$ for 2 groups: $\mathcal{S}_{\text{gold human}}$ and $\mathcal{S}_{\text{gold machine}}$. Intuitively, a value such as $\bar{r}_{\text{gold machine}}(c) = 0.01$ means that, on average, *1%* of all retained patches in a machine-labeled sample belong to a cluster c; a difference of 0.15 corresponds to a 15% shift in the average patch share.

Motif discriminativeness across datasets and proxy models. Given gold-labeled groups $\mathcal{S}_{\text{gold human}}$ and $\mathcal{S}_{\text{gold machine}}$, we quantify how strongly a motif cluster separates classes via the difference in mean motif rates

$$\Delta r(c) \;=\; \bar{r}_{\text{gold machine}}(c) \;-\; \bar{r}_{\text{gold human}}(c). \tag{19}$$

Using the aggregate statistics (top-3 clusters by $|\Delta r(c)|$ per setting), we find that the most discriminative motif often shifts by several percentage points in average patch share, and in some cases by more than 10 points. Table 10 reports the top-1 cluster per dataset/proxy/preprocessing variant, including bootstrap confidence intervals for $\Delta r(c)$ and a permutation-test p-value. In HC3, the strongest motif (cid=24) is *machine-enriched*, increasing from 9.7% (gold human) to 20.7% (gold machine), $\Delta r = +11.0$ points with a tight 95% CI $[9.3, 12.7]$. In contrast, for RAID the top motifs are consistently *human-enriched* under all tested proxy models and both preprocessing variants (e.g., $\Delta r = -10.5$ points for Cohere under the unified variant). Qualitatively, the corresponding example patches (Fig. 2i) show similar, repeatable local structures: for human-leaning patch clusters, "islands" and isolated flashes of attention are observed across all datasets and models, while the only identified heavy machine-inclined cluster of HC3/GPT-neo exhibits a prevalence of the horizontal bands. The notable outlier is GPT-neo for RAID, which is explained by HDBSCAN producing a very low amount of dense clusters (n = 4), resulting in a high intra-class variance. Such observations support the interpretation of clusters as visually coherent "motif families".

Patch-wise saliency and ablation are computed per retained patch. To relate motif *prevalence* to motif *importance*, we assign each retained patch an (i) Grad-CAM(Selvaraju et al., 2019) score and (ii) a zeroing-ablation score, and then aggregate these scores by cluster. Concretely, for each retained patch index $(k, u, v) \in \mathcal{I}_s$ we compute: (i) a Grad-CAM activation $g_{s,k,u,v}$ at the last convolutional stage (the same 16×16 grid that defines the patches), and (ii) an ablation-induced logit change $\delta_{s,k,u,v}$ obtained by zeroing the corresponding 8×8 region in the input block A_k and re-evaluating the detector. We then summarize cluster-level importance by averaging over all occurrences assigned to cluster c:

$$\begin{aligned} \bar{g}(c) &= \mathbb{E}[g_{s,k,u,v} \mid c_{k,u,v} = c]\,, \\ \bar{\delta}(c) &= \mathbb{E}[\delta_{s,k,u,v} \mid c_{k,u,v} = c]\,. \end{aligned} \tag{20}$$

Here $\bar{\delta}(c)$ captures the *average marginal effect* of removing a single motif instance, while $\bar{r}_g(c)$ captures *how frequently* the motif occurs in a given group.

Why prevalence need not correlate with saliency or ablation. Empirically, we observe that clusters with the largest $|\Delta r(c)|$ are not necessarily the clusters with the highest $\bar{g}(c)$ or $|\bar{\delta}(c)|$. Aggregated correlations across all datasets and models are weak and unstable: Across top-15 clusters for Grad-CAM Pearson $\rho =$ 0.143$\pm$0.23, Spearman $\rho =$ 0.10$\pm$0.29; for Zero-ablation Pearson $\rho =$ -0.03$\pm$0.36, Spearman $\rho =$ -0.09$\pm$0.25. Across the top-3 clusters, correlation becomes moderate for GRAD-CAM but still unstable: Pearson $\rho = 0.373 \pm 0.66$, Spearman $\rho = 0.37 \pm 0.54$; while it slightly improves for Zero-ablation, Pearson $\rho = 0.09 \pm 0.67$, Spearman $\rho = 0.12 \pm 0.64$. We attribute this mismatch to the following reasons: First, a dataset imbalance can decouple prevalence from importance. Taking into account the significant imbalance of used datasets, the RAID subset and HC3, training under skewed prior biases the detector toward features that optimize empirical risk for the majority class, so a motif may be strongly pronounced in $\Delta r(c)$ yet have muted average Grad-CAM/ablation effects. Second, for Grad-CAM and zeroing-ablation the same motif family may be decisive only in certain positions or alongside other motifs; averaging $\bar{g}(c)$ and $\bar{\delta}(c)$ across all occurrences can weaken these conditional effects.

Taken together, these findings suggest that motif analysis is best interpreted as a complementary view: $\Delta r(c)$ reveals dataset and proxy-model dependent differences in local attention map structure, while $\bar{g}(c)$ and $\bar{\delta}(c)$ reflect how the trained detector *uses* (or ignores) individual motif instances at decision time conditioned on the training data.

Implications. Overall, the presence of statistically reliable shifts in patch motif rates between gold machine and gold human groups (Table 10) supports our hypothesis that the internal dynamics of a proxy G_θ induce detectable structure in attention-based attribution maps beyond surface-level text statistics. At the same time, the weak alignment between prevalence and saliency cautions against interpreting frequent motifs as predictions of the model behaviour, instead, they appear to function as stable, repeatable signatures that the detector can exploit in combination with other cues.

6 Conclusion

We presented AEyeDE, an attribution-based framework for AI-generated text detection that leverages attention-derived attribution maps from a proxy Transformer as a signal to learn a model to distinguish between human-written and AI-generated text. Across both encoder-decoder translation benchmarks (WMT14, UN) and decoder-only generation datasets (HC3, RAID), as well as adversarial attacks and unseen cross-dataset samples, AEyeDE achieves competitive performance and shows strong results. This is achieved by training a lightweight CNN model on a relatively small amount of data.

Beyond evaluation metrics, we provide an analysis of localized attribution patterns and show that they systematically differ between human and AI-generated text. Overall, our results suggest that internal attribution behavior offers a complementary and effective signal for reliable authorship detection, motivating further work on broader robustness settings and alternative attribution sources.

Limitations Our study has some limitations. First, the proposed framework assumes *white-box* access to a proxy Transformer model in order to extract attention-based attribution maps. While this assumption may limit applicability in fully black-box settings, our experiments indicate that attribution patterns generalize across generator families, suggesting that exact access to the true generator is not strictly required.

Furthermore, our implementation focuses on attention-based attributions, which offer a favorable trade-off between informativeness and computational cost for large models and long sequences. Investigating alternative attribution methods, such as gradient-based saliency, may further enrich the analysis, but is left for future work due to their higher computational overhead. Additionally, we computed attention attribution by averaging over all layers and heads. Investigating a more fine-grained selection strategy for this task remains for future work.

References

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah, et al. The arageneval shared task on arabic authorship style transfer and ai generated text detection. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 1–13, 2025.

Hyeseon Ahn, Shinwoo Park, Suyeon Woo, and Yo-Sub Han. Ditto: A spoofing attack framework on watermarked llms via knowledge distillation. *arXiv preprint arXiv:2510.10987*, 2025.

Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros (eds.). *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE, January 2025. International Conference on Computational Linguistics. URL https://aclanthology.org/2025.genaidetect-1.0/.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. Detection of human and machine-authored fake news in Urdu. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3419–3428, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.170/.

Ekaterina Artemova, Jason S Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. Beemo: Benchmark of expert-edited machine-generated outputs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6992–7018, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.357. URL https://aclanthology.org/2025.naacl-long.357/.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. Beyond easy wins: A text hardness-aware benchmark for llm-generated text detection. *arXiv preprint arXiv:2507.15286*, 2025.

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. *arXiv preprint arXiv:2310.05130*, 2023.

Kyle Bittle and Omar El-Gayar. Generative ai and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4):296, 2025.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi

Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Joseph Cornelius, Oscar Lithgow-Serrano, Sandra Mitrović, Ljiljana Dolamic, and Fabio Rinaldi. Bust: Benchmark for the evaluation of detectors of llm-generated text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8029–8057, 2024.

Xinyue Cui, Johnny Wei, Swabha Swayamdipta, and Robin Jia. Robust data watermarking in language models by injecting fictitious knowledge. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14292–14306, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12763–12771, 2023.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

Markus Frohmann, Gabriel Meseguer-Brocal, Markus Schedl, and Elena V. Epure. Double entendre: Robust audio-based AI-generated lyrics detection via multi-view fusion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1914–1926, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.98/.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. XDAC: XAI-driven detection and attribution of LLM-generated news comments in Korean. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22728–22750, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1108. URL https://aclanthology.org/2025.acl-long.1108/.

Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. Xdac: Xai-driven detection and attribution of llm-generated news comments in korean. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22728–22750, 2025b.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain

Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla,

Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

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, 2024. URL https://arxiv.org/abs/2401.12070.

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

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095, 2023.

Beining Huang, Du Su, Fei Sun, Qi Cao, Huawei Shen, and Xueqi Cheng. Low-entropy watermark detection via bayes' rule derived detector. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14330–14344, 2025a.

He Huang, Nan Sun, Massimiliano Tani, Yu Zhang, Jiaojiao Jiang, and Sanjay Jha. Can llm-generated misinformation be detected: A study on cyber threat intelligence. *Future Generation Computer Systems*, 173:107877, 2025b. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2025.107877. URL https://www.sciencedirect.com/science/article/pii/S0167739X25001724.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808–1822, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.164. URL https://aclanthology.org/2020.acl-main.164/.

Michael Ivanitskiy, Cecilia Diniz Behn, and Samy Wu Fung. Motifs in attention patterns of large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8971–8983, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.469/.

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*, pp. 17061–17084. PMLR, 2023.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. RankGen: Improving text generation with large ranking models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022*

Conference on Empirical Methods in Natural Language Processing, pp. 199–232, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.15. URL https://aclanthology.org/2022.emnlp-main.15/.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575c450013d0e99e4b0ecf82bd1afaa4-Paper-Conference.pdf.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Feature-level insights into artificial text detection with sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25727–25748, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.1321/.

Salima Lamsiyah, Saad Ezzini, Abdelkader El Mahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. M-daigt: A shared task on multi-domain detection of ai-generated text. *arXiv preprint arXiv:2511.11340*, 2025.

Jiatao Li and Xiaojun Wan. Who writes what: Unveiling the impact of author roles on AI-generated text detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26620–26658, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1292/.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3091–3113, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.155/.

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, 2024a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. On the detectability of chatgpt content: Benchmarking, methodology, and evaluation through the lens of academic writing. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2236–2250, 2024b.

Xinyang Lu, Jingtan Wang, Zitong Zhao, Zhongxiang Dai, Chuan-Sheng Foo, See Kiong Ng, and Bryan Kian Hsiang Low. Wasa: Watermark-based source attribution for large language model-generated data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23791–23824, 2025.

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. An entropy-based text watermarking detection method. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11724–11735, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.630. URL https://aclanthology.org/2024.acl-long.630/.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In Wanxiang Che, Joyce Nabende, Ekaterina

Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 727–752, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.36/.

Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. Watermarking large language models: An unbiased and low-risk method. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7939–7960, 2025.

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*, pp. 24950–24962. PMLR, 2023.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Georg Niess and Roman Kern. Ensemble watermarks for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2903–2916, 2025.

OpenAI. Chatgpt. https://chat.openai.com, 2023. Large language model.

Andrea Pedrotti, Michele Papucci, Cristiano Ciaccio, Alessio Miaschi, Giovanni Puccetti, Felice Dell'Orletta, and Andrea Esuli. Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3010–3031, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.156/.

Xinlin Peng, Ying Zhou, Ben He, Le Sun, and Yingfei Sun. Hidding the ghostwriters: An adversarial evaluation of ai-generated student essay detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10406–10419, 2023.

Jiawen Qi, Chang Gao, Zhaochun Ren, and Qinyu Chen. Deltallm: A training-free framework exploiting temporal sparsity for efficient edge llm inference. *arXiv preprint arXiv:2507.19608*, 2025.

Rafael Alberto Rivera Soto, Barry Y. Chen, and Nicholas Andrews. Mitigating paraphrase attacks on machine-text detection via paraphrase inversion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4421–4433, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.227/.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. Inseq: An interpretability toolkit for sequence generation models. In Danushka Bollegala, Ruihong Huang, and Alan Ritter (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 421–435, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-demo.40. URL https://aclanthology.org/2023.acl-demo.40/.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge, Xing Yang, and Yanghe Feng. Textdefense: Adversarial text detection based on word importance entropy. *arXiv preprint arXiv:2302.05892*, 2023.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/2023.findings-emnlp.827/.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22015–22036, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1069/.

Vasiliki Tassopoulou, George Retsinas, and Petros Maragos. Enhancing handwritten text recognition with n-gram sequence decomposition and multitask learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10555–10560. IEEE, 2021.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.61/.

Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, and Maksim Kuprashevich. Llmtrace: A corpus for classification and fine-grained localization of ai-written text. *arXiv preprint arXiv:2509.21269*, 2025.

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. *arXiv preprint arXiv:2109.13296*, 2021.

Ashok Urlana, Aditya Saibewar, Bala Mallikarjunarao Garlapati, Charaka Vinayak Kumar, Ajeet Singh, and Srinivasa Rao Chalamala. TrustAI at SemEval-2024 task 8: A comprehensive analysis of multi-domain machine generated text detection techniques. In Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (eds.), *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 927–934, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.semeval-1.134. URL https://aclanthology.org/2024.semeval-1.134/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (eds.), *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 2057–2079, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.semeval-1.279. URL https://aclanthology.org/2024.semeval-1.279/.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1369–1407, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.83. URL https://aclanthology.org/2024.eacl-long.83/.

Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. MorphMark: Flexible adaptive watermarking for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4842–4860, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.240. URL https://aclanthology.org/2025.acl-long.240/.

Junchao Wu, Runzhe Zhan, Derek Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 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*, 51 (1):275–338, March 2025. doi: 10.1162/coli_a_00549. URL https://aclanthology.org/2025.cl-1.8/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12298–12319, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.601. URL https://aclanthology.org/2025.acl-long.601/.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1561/.