Beyond Global Scalars: Synergizing Token-Level Statistics and Deep Semantics for Adversarial AIGC Text Detection

Peiming Li[1,2] , **Yifan Wang**[1] , **Zhiyuan Hu**[1,2] , **Shiyu Li**[1] ,
Zheng Wei[1,*], **Yang Tang**[1,*,†]
[1]Tencent BAC
[2]School of Electronic and Computer Engineering, Peking University
hemingwei@tencent.com, ethanntang@tencent.com

Homepage | Code | MOSAIC Dataset

Abstract

The rapid evolution of large language models necessitates robust machine-generated text detection. Existing paradigms typically follow two isolated tracks. Training-free methods rely on global statistical scalars such as perplexity, while training-based methods utilize semantic hidden states. Both approaches exhibit fundamental vulnerabilities in adversarial scenarios. Global scalars act as lossy compressions that obscure local probabilistic burstiness in interleaved texts, whereas pure semantic models overfit to specific fingerprints and remain susceptible to spoofing. To expose these flaws, we introduce **MOSAIC**, a comprehensive adversarial benchmark comprising 16000 samples across a full-granularity attack spectrum. To address these challenges, we propose **NeuroStat**, an end-to-end framework bridging the statistical and semantic gap. NeuroStat captures uncompressed token-level probabilistic logits alongside deep semantic hidden states from a single causal language model backbone. We fuse these heterogeneous signals through Macro-State Residual Modulation, which adaptively calibrates local convolutional features using global uncertainty indicators. Orthogonal and contrastive losses further ensure the learning of complementary representations. Extensive experiments demonstrate that NeuroStat maintains exceptional robustness on MOSAIC compared to the severe degradation of state-of-the-art methods, establishing a new standard for adversarial text detection. Code and the MOSAIC benchmark are available at https://github.com/TencentBAC/NeuroStat.

*Corresponding authors.
†Project Lead.

1 Introduction

The proliferation of advanced Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; OpenAI et al., 2024; GLM-5-Team et al., 2026; Team et al., 2026) has blurred the boundary between human-written and machine-generated text (Clark et al., 2021). While this technological leap empowers productivity, it exacerbates the spread of misinformation (Zellers et al., 2019). Consequently, Machine-Generated Text Detection (MGTD) has emerged as a critical research frontier.

Current MGTD methodologies predominantly follow two isolated paradigms, Training-Free (TF) methods that calculate global statistical scalars (e.g., perplexity) (Mitchell et al., 2023; Bao et al., 2024b), and Training-Based (TB) methods that fine-tune black-box encoders to capture semantic styles (Solaiman et al., 2019; Verma et al., 2024). Recent hybrid strategies (Chen et al., 2024; Fu et al., 2025) attempt to optimize scoring models but still rely on lossy global scalars. In real-world applications, malicious users rarely generate entire documents directly. Instead, they employ Interleaved Human-AI Writing or Statistical Hijacking (Sadasivan et al., 2025; Krishna et al., 2023). As illustrated in Fig. 1, existing paradigms are easily bypassed, TF methods average out local probability bursts and are spoofed by high-entropy instructions (Fig. 1a), while TB methods overfit to specific semantic fingerprints, leaving them susceptible to style manipulation (Fig. 1b).

To systematically diagnose these vulnerabilities, we introduce **MOSAIC** (**M**ultidimensional **O**bfuscation and **S**poofing **A**dversarial **I**nterleaved **C**orpus), the most comprehensive adversarial MGTD benchmark to date. Unlike existing datasets (Guo et al., 2023; Dugan et al., 2024), MO-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50091–50106
October 24-29, 2026 ©2026 Association for Computational Linguistics

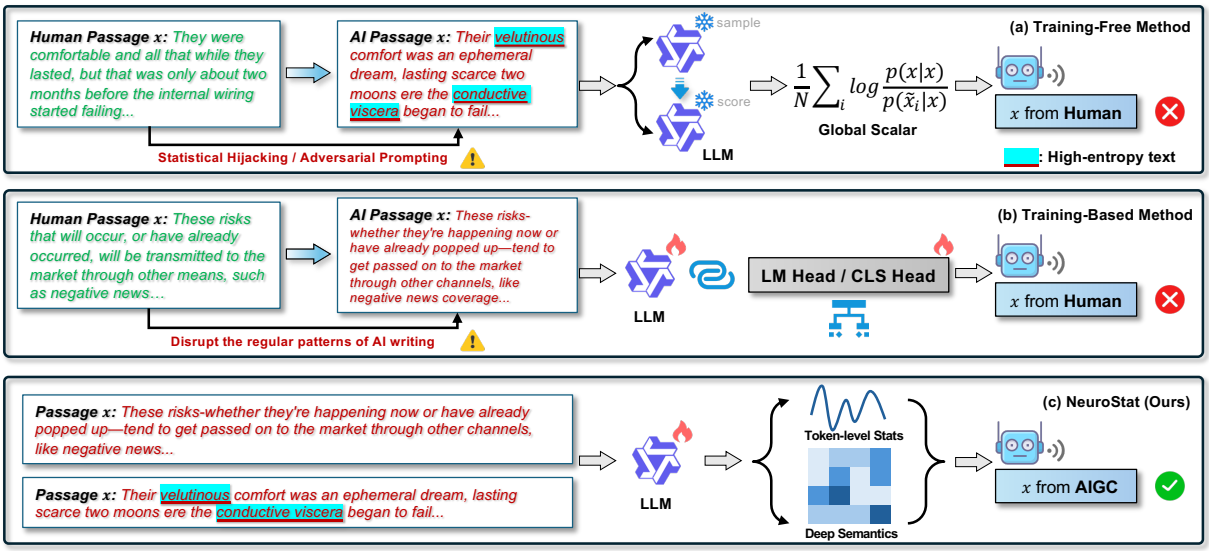

Figure 1: **Comparison of detection paradigms under adversarial attacks.** Existing methods fail due to their reliance on either (a) easily hijacked global statistical scalars (Training-Free) or (b) overfitted semantic patterns that can be disrupted by style manipulation (Training-Based). (c) NeuroStat robustly detects adversarial AIGC by synergizing uncompressed token-level statistical trajectories and deep semantic representations.

SAIC constructs a full-granularity attack spectrum encompassing 8 categories and 36 sub-methods, employing a cross-allocation mechanism across six frontier LLMs to mitigate model fingerprint bias.

To overcome the inherent flaws exposed by MOSAIC, we propose **NeuroStat** (**Neu**ral Embeddings and **R**epresentations **O**rthogonalized with **Sta**tistical **T**rajectories), a novel end-to-end framework that bridges the TF-TB methods' gap. We argue that TF and TB provide orthogonal but complementary views, TF captures the probabilistic mechanics of decoding, while TB captures contextual semantics. NeuroStat extracts both token-level statistical trajectories and deep semantic representations from a single CausalLM backbone.

Crucially, instead of compressing probabilities into a vulnerable scalar, NeuroStat preserves the full sequential trajectory. To fuse these heterogeneous signals, we design a Macro-State Residual Modulation (MSRM) mechanism. MSRM utilizes global uncertainty indicators to dynamically amplify local probability anomalies, effectively capturing hidden AI bursts within fluent human contexts. Trained with orthogonal and contrastive objectives, NeuroStat learns discriminative and non-redundant representations, maintaining exceptional robustness against adversarial attacks (Fig. 1c).

Our contributions are summarized as follows:

- **A Paradigm Shift in Detection Architecture.** We propose NeuroStat, the first framework to achieve end-to-end fusion of probabilistic trajectories and semantic artifacts. By replacing lossy global scalars with MSRM, it overcomes dilution and spoofing flaws.
- **A Full-Spectrum Adversarial Benchmark:** We construct MOSAIC, a benchmark featuring an unprecedented 8×36 attack matrix and multi-model cross-generation, establishing a new standard for MGTD robustness.
- **State-of-the-Art Robustness.** Extensive experiments demonstrate that NeuroStat significantly outperforms existing baselines across diverse domains, models, and attack granularities, maintaining high accuracy even in extreme adversarial scenarios.

2 Related Work

Training-Free (TF) Detection Methods. TF methods distinguish machine-generated text by leveraging LLMs' inherent probabilistic mechanics via zero-shot metrics, including log-likelihood, entropy (Lavergne et al., 2008; Solaiman et al., 2019; Gehrmann et al., 2019), probability curvature (Mitchell et al., 2023; Bao et al., 2024b), log-rank (Su et al., 2023), n-gram divergence (Yang et al., 2023), cross-perplexity (Hans et al., 2024), and distribution recovery (Bao et al., 2024a). However, these methods universally compress sequential token-level dynamics into a lossy global scalar,

making them highly vulnerable to dilution in interleaved human-AI texts and statistical hijacking.

Training-Based (TB) Detection Methods. TB methods treat detection as a supervised task to capture semantic artifacts, employing encoder fine-tuning (Liu et al., 2019; Zellers et al., 2019; Ippolito et al., 2020), adversarial training (Hu et al., 2023; Verma et al., 2024; Wu et al., 2023), or alignment-based optimization (Chen et al., 2024; Fu et al., 2025). Despite near-perfect in-domain accuracy, TB methods inherently overfit to specific semantic fingerprints. Diverging from paradigms relying on isolated global scalars or black-box semantics, NeuroStat synergizes uncompressed probabilistic trajectories with semantic representations.

Detection Benchmarks and Adversarial Robustness. Early detection benchmarks focused on basic generation (Uchendu et al., 2021; Guo et al., 2023) before expanding to multi-domain settings (Wang et al., 2024; Macko et al., 2023; Li et al., 2024). Concurrently, studies reveal detectors are easily bypassed by paraphrasing, homoglyphs, and prompt engineering (Sadasivan et al., 2025; Krishna et al., 2023; Gagiano et al., 2021; Lu et al., 2024). While RAID (Dugan et al., 2024) standardized robustness evaluation with 11 attacks, existing benchmarks still suffer from limited attack surfaces and model bias. To our knowledge, MOSAIC is the first to provide a full-granularity taxonomy of 36 attacks coupled with multi-model cross-generation.

3 Methodology

We present **NeuroStat**, an end-to-end dual-branch framework that bridges the Training-Free (TF) and Training-Based (TB) paradigms for robust machine-generated text detection (illustrated in Fig. 2). The core philosophy of NeuroStat is to decouple the detection process into two orthogonal perspectives: the probabilistic mechanics of the decoding process (TF) and the contextual semantic artifacts (TB). Given an input token sequence $\mathbf{x} = (x_1, \ldots, x_T)$ of length T, a single forward pass through the backbone $\mathcal{M}$ yields the full vocabulary logit matrix $\mathbf{L} \in \mathbb{R}^{T \times |\mathcal{V}|}$ and the last-layer hidden states $\mathbf{H} \in \mathbb{R}^{T \times d}$, where $|\mathcal{V}|$ is the vocabulary size and d is the hidden dimension. These heterogeneous signals are routed to their respective branches, enabling synergistic feature extraction without requiring multiple model invocations.

3.1 Probabilistic Trajectories Branch (TF)

Existing TF methods typically compress token probabilities into a lossy global scalar (e.g., perplexity), which inadvertently smooths out local anomalies. Instead, our TF branch is designed to preserve the complete sequential structure of the probability landscape. Specifically, we extract the shifted log-probability trajectory $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{T-1})$ of the actual tokens, where:

$$\ell_t = \log \big(\text{softmax}(\mathbf{L}_t)\big)_{x_{t+1}}. \tag{1}$$

Here, $\mathbf{L}_t$ denotes the logit vector at step t, and x_{t+1} is the actual next token in the sequence.

Unlike machine-generated texts that maintain consistently high probabilities, human-authored content inherently manifests pronounced variance in $\boldsymbol{\ell}$ (i.e., local burstiness). To capture these localized dynamics and identify abrupt likelihood decays indicative of human-AI boundaries, we process $\boldsymbol{\ell}$ through a hierarchical 1D-CNN:

$$\mathbf{c} = f_{\text{CNN}}(\boldsymbol{\ell}) \in \mathbb{R}^{d_c}. \tag{2}$$

This architecture functions as a localized n-gram anomaly detector, effectively extracting multi-scale probabilistic deviations while preserving strict positional awareness.

Macro-State Residual Modulation (MSRM). While the 1D-CNN effectively extracts localized dynamics, it lacks global contextual awareness. An abrupt probability decay could signify a naturally rare word in human writing or a hallucination boundary in AI generation. To disambiguate such signals, MSRM conditions the CNN features on two macro-state uncertainty indicators: mean entropy $\bar{E}$ and mean log-rank $\bar{R}$:

$$\bar{E} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathcal{H}\big(\text{softmax}(\mathbf{L}_t)\big), \tag{3}$$

$$\bar{R} = \frac{1}{T-1} \sum_{t=1}^{T-1} \log \text{rank}_t(x_{t+1}), \tag{4}$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy, and $\text{rank}_t(x_{t+1})$ is the descending rank of the target token at step t. These scalars, concatenated as $\mathbf{s} = [\bar{E}; \bar{R}] \in \mathbb{R}^2$, quantify the model's sequence-level confidence. This design formalizes an asymmetry long exploited by training-free detectors, where global confidence and local rank or probability deviations jointly determine separability rather than either signal alone (Gehrmann et al., 2019;

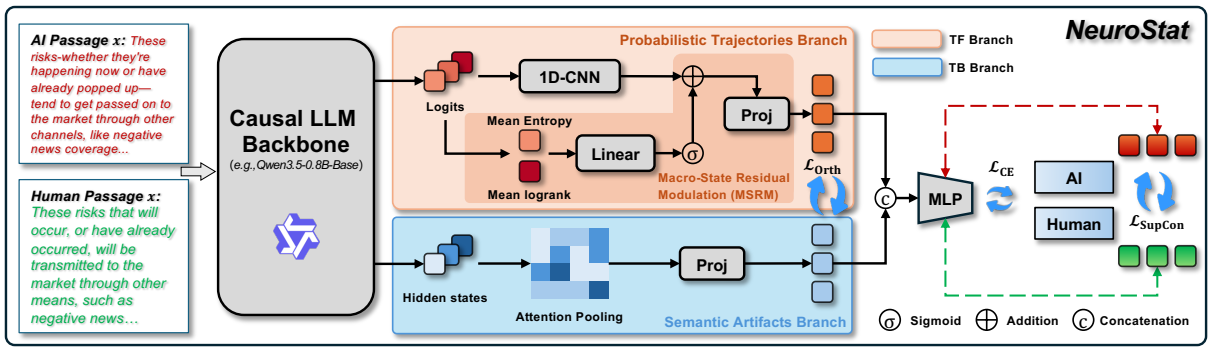

Figure 2: **The overall architecture of NeuroStat.** A CausalLM backbone simultaneously extracts token logits and hidden states. The **Probabilistic Trajectories Branch** (top) captures local statistical patterns, which are dynamically calibrated by the **MSRM** mechanism using global uncertainty indicators. Concurrently, the **Semantic Artifacts Branch** (bottom) extracts contextual styles via attention pooling. The fused representations are jointly optimized with $\mathcal{L}_{\text{Orth}}$ and $\mathcal{L}_{\text{SupCon}}$ objectives to ensure complementary and robust feature learning.

Mitchell et al., 2023; Hans et al., 2024). They modulate the CNN features via a residual gating mechanism:

$$\tilde{\mathbf{c}} = \mathbf{c} \odot (1 + \sigma(\mathbf{W}_g \mathbf{s})), \tag{5}$$

where $\mathbf{W}_g \in \mathbb{R}^{d_c \times 2}$ is a learnable projection and σ is the sigmoid function. Conceptually, when the global entropy is low, reflecting the high overall confidence typical of AI generation, MSRM acts as an adaptive amplifier for localized anomaly signals. The modulated features are subsequently projected to yield the final TF representation $\mathbf{z}_{\text{tf}} \in \mathbb{R}^{d_z}$. We empirically verify this design by correlating the learned gate activation with the global entropy $\bar{E}$ on held-out samples, obtaining a Pearson correlation of -0.62. This confirms that the network learns to amplify local features precisely when global confidence is high, without this behavior being explicitly imposed.

3.2 Semantic Artifacts Branch (TB)

While the TF branch focuses on how the text was generated, the TB branch focuses on what was generated. Machine-generated texts often contain specific stylistic markers, repetitive phrasing patterns, or rigid discourse structures (e.g., "delve into", "in conclusion"). To capture these high-level semantic artifacts from the hidden states $\mathbf{H}$, we employ a learnable attention pooling mechanism:

$$\alpha_t = \frac{\exp(\mathbf{w}_a^\top \mathbf{h}_t)}{\sum_{j=1}^{T} \exp(\mathbf{w}_a^\top \mathbf{h}_j)}, \quad \mathbf{h}_{\text{pool}} = \sum_{t=1}^{T} \alpha_t \, \mathbf{h}_t, \tag{6}$$

where $\mathbf{w}_a \in \mathbb{R}^d$ is a learnable query vector. Unlike relying solely on the last token, this attention mechanism enables the model to dynamically assign higher weights to the most semantically discriminative tokens, regardless of sequence length. The pooled representation is then projected to yield the TB representation $\mathbf{z}_{\text{tb}} \in \mathbb{R}^{d_z}$.

3.3 Training Objectives

The representations from both branches are concatenated and passed through a Multi-Layer Perceptron (MLP) to produce the binary classification logits $\hat{\mathbf{y}} = \text{MLP}([\mathbf{z}_{\text{tf}}; \mathbf{z}_{\text{tb}}])$. To ensure that NeuroStat learns robust and non-redundant features, it is optimized with three complementary objectives:

Cross-Entropy Loss ($\mathcal{L}_{CE}$). The primary classification loss optimized over the predicted probability of the AIGC class.

Supervised Contrastive Loss ($\mathcal{L}_{\text{SupCon}}$). To enhance the model's robustness against out-of-distribution LLMs, we apply $\mathcal{L}_{\text{SupCon}}$ to the ℓ_2-normalized fused vector $\mathbf{e}_i$:

$$\mathcal{L}_{\text{SupCon}} = \frac{-1}{B} \sum_{i=1}^{B} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\mathbf{e}_i \cdot \mathbf{e}_j / \tau)}{\sum_{k \neq i} \exp(\mathbf{e}_i \cdot \mathbf{e}_k / \tau)}, \tag{7}$$

where $P(i)$ contains same-class samples in batch B, and τ is the temperature. This pulls same-class representations together while pushing apart different classes, improving robustness against distributional shifts.

Orthogonal Penalty ($\mathcal{L}_{\text{Orth}}$). A critical challenge in dual-branch architectures is feature collapse, where one branch dominates or both learn redundant information. Assuming both branches are projected to the same dimension d_z, we force the

TF and TB branches to span orthogonal subspaces by minimizing their squared cosine similarity:

$$\mathcal{L}_{\text{Orth}} = \frac{1}{B}\sum_{i=1}^{B}\left(\frac{\mathbf{z}_{\text{tf}}^{i}\cdot\mathbf{z}_{\text{tb}}^{i}}{\|\mathbf{z}_{\text{tf}}^{i}\|\|\mathbf{z}_{\text{tb}}^{i}\|}\right)^{2}. \quad (8)$$

This strict regularization ensures that the semantic branch does not redundantly encode statistical patterns, and vice versa. Measured on held-out samples, the mean squared cosine similarity between the two branches drops from 0.358 without $\mathcal{L}_{\text{Orth}}$ to 0.018 with $\mathcal{L}_{\text{Orth}}$, closely approaching the 0.016 floor observed for independent random unit vectors of the same dimensionality. This indicates that the two branches become effectively independent rather than merely less correlated.

Total Objective. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{\text{SupCon}}\mathcal{L}_{\text{SupCon}} + \lambda_{\text{Orth}}\mathcal{L}_{\text{Orth}}, \quad (9)$$

where λ_{SupCon} and λ_{Orth} are hyperparameters balancing the auxiliary losses.

4 The MOSAIC Benchmark

To systematically diagnose the vulnerabilities of existing MGTD paradigms, we introduce **MOSAIC**, a large-scale adversarial benchmark comprising 16000 human-AI text pairs. MOSAIC is generated through a rigorous pipeline emphasizing multi-granularity attacks and cross-model generation.

4.1 Source Data Collection and Purification

To ensure domain diversity, we collect 36753 raw human-written texts from eight distinct sources. To obtain high-confidence human-written seed texts, we implement a stringent four-stage purification pipeline: (1) **Exact Deduplication**; (2) **Rule-Based Filtering** to remove truncated, HTML-heavy, or non-English texts; (3) **Near-Duplicate Removal** via 64-bit SimHash and Locality-Sensitive Hashing (LSH); and (4) **Multi-Dimensional Quality Scoring**, evaluating Log-TTR, punctuation, and syntax, retaining texts with scores ≥ 85. This rigorous process yields 16000 high-quality human seed texts.

4.2 Full-Granularity Adversarial Taxonomy

As summarized in Tab. 1, existing benchmarks cover limited attack surfaces. Even comprehensive datasets like RAID and MIRAGE encompass at most 11 attacks across 2 categories, leaving significant blind spots. In contrast, MOSAIC introduces an unprecedented taxonomy of 36 distinct adversarial prompts across 8 linguistic granularities, achieving full-spectrum coverage:

- **Interleaved Human-AI Text:** Paragraph-level content replacement and sentence-level mixed interleaving attacks.
- **Statistical Hijacking:** Manipulating AI statistical signatures (e.g., perplexity) via controlled synonym swapping.
- **Multi-hop Translation Laundering:** Erasing latent AI fingerprints via cyclic multilingual translation chains.
- **Paraphrase & Polish:** A continuous rewriting spectrum from light polishing to deep structural paraphrasing.
- **Prompt Injection & Role-playing:** Altering default semantic styles using persona adoption and explicit anti-detection instructions.
- **Character & Encoding Level:** Injecting adversarial homoglyphs and invisible zero-width characters to disrupt tokenizers.
- **Recursive & Multi-Pass:** Iteratively refining generated texts across multiple models to smooth artifacts.
- **Domain-Specific & Format:** Forcing highly specialized writing formats (e.g., academic hedging, internet slang).

4.3 Multi-Model Cross-Generation

Previous datasets over-rely on a single model family, causing TB detectors to overfit to specific "semantic fingerprints". To mitigate generator-specific fingerprint bias, MOSAIC employs a random cross-allocation mechanism across six frontier LLMs: GLM-5.0, GPT-5.4, Gemini-3.1-Pro, MiniMax-M2.7, Claude-Opus-4-6, and Kimi-K2.5. Specifically, the 16000 seed texts are stratified by length. Within each stratum, the 36 adversarial prompts are uniformly assigned, and each prompt-seed pair is allocated to one of the six LLMs. This model-agnostic design forces detectors evaluated on MOSAIC to learn the intrinsic boundary between human and AI texts, rather than merely memorizing the stylistic quirks of a specific generator.

5 Experiments

5.1 Experiment Settings

Datasets and Tasks. To comprehensively evaluate NeuroStat, we conduct experiments on three bench-

Benchmark	Data Statistics		MGT Tasks			Adversarial Attack Coverage								Summary	
	Size	#Dom.	Gen.	Pol.	Rew.	Inter-leaved	Stat. Hijack	Trans-lation	Para-phrase	Prompt Inj.	Char./ Enc.	Recur-sive	Dom./ Fmt.	#Cat.	#Atk.
TuringBench (Uchendu et al., 2021)	168.6K	1	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0	0
HC3 (Guo et al., 2023)	85K	3	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0	0
M4 (Wang et al., 2024)	24.5K	4	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	1	1
MAGE (Li et al., 2024)	29K	5	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	2	2
RAID (Dugan et al., 2024)	628.7K	4	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗	2	11
DetectRL (Wu et al., 2025)	134.4K	4	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	1	3
HART (Bao et al., 2025)	84K	4	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	1	3
ImBD (Chen et al., 2024)	10K	3	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	2	2
MIRAGE (Fu et al., 2025)	93.8K	5	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	2	3
MOSAIC (Ours)	**16K**	**5**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**8**	**36**

Table 1: Comparison with existing MGT detection benchmarks. Our benchmark provides **36** sub-methods across **8** linguistic granularity levels with full category coverage, generated via uniform random model assignment over 6 frontier LLMs. ✓ = supported; ✗ = not supported. Comprehensive details are provided in the Appendix Sec. B

marks, each targeting a different capability of the detector. **MIRAGE** evaluates *generalization* across three tasks (GENERATE, POLISH, and REWRITE) under both Disjoint-Input (DIG) and Shared-Input (SIG) settings. **MOSAIC**, our proposed benchmark, evaluates *adversarial robustness* through 36 fine-grained attacks spanning 8 categories. **ImBD Test Set** evaluates *standard detection* on human-written texts and their machine-polished counterparts across three domains. The corresponding results are provided in the Appendix Sec. C.

Baselines. We compare NeuroStat against 11 state-of-the-art MGTD methods, including **Training-Free (TF) methods**: Likelihood (Solaiman et al., 2019), LogRank (Ippolito et al., 2020), Entropy (Gehrmann et al., 2019), LRR (Su et al., 2023), DNA-GPT (Yang et al., 2023), NPR (Su et al., 2023), DetectGPT (Mitchell et al., 2023), and Fast-DetectGPT (Bao et al., 2024b); and **Training-Based (TB) methods**: RoBERTa-Base/Large (Liu et al., 2019), ImBD (Chen et al., 2024) and DetectAnyLLM (Fu et al., 2025). On the adversarial MOSAIC benchmark (Tab. 3), we additionally report Binoculars (Hans et al., 2024) and TextFluoroscopy (Yu et al., 2024) as paraphrase-robust baselines, while omitting NPR and DetectGPT, whose perturbation-based scoring degrades to near-random performance under heavy adversarial rewriting. All TF baselines share the same Qwen2-0.5B reference model, and all TB baselines are retrained on the identical 500-pair split for a fair comparison.

Implementation Details. **(1) Base Model:** We instantiate NeuroStat using two lightweight CausalLM backbones: Qwen2-0.5B and Qwen3.5-0.8B. **(2) Training Paradigm:** To ensure a strictly fair comparison, NeuroStat is trained on only 500 pairs of human-written and machine-polished texts, identical to the ImBD training split. Diverging from previous methods that freeze the backbone or rely on parameter-efficient fine-tuning (e.g., LoRA), the entire NeuroStat framework, including the CausalLM backbone and the dual-branch modules, is trained end-to-end. This 500-pair budget is chosen purely as a fairness constraint rather than an assumed sufficiency threshold, and a training-size sensitivity study is provided in Appendix Sec. E.

5.2 Main Results

Generalization on MIRAGE Benchmark. Tab. 2 presents the comprehensive results on the MIRAGE benchmark. Existing baselines exhibit significant performance degradation when faced with diverse domains and tasks. In contrast, NeuroStat demonstrates exceptional generalization capabilities. Under the challenging MIRAGE-SIG setting, NeuroStat (Qwen3.5-0.8B) achieves an AUROC of **0.9954**, **0.9739**, and **0.9665** across the three tasks. This strong performance with only 500 training samples highlights the superiority of our dual-branch feature extraction in capturing universal AI artifacts across diverse generative distributions.

Robustness on MOSAIC Benchmark. NeuroStat's superiority is most pronounced under adversarial conditions (Tab. 3). As hypothesized, existing paradigms suffer catastrophic failures against targeted attacks. TF methods like Fast-DetectGPT plummet under *Statistical Hijacking* (34.5) and *Interleaved* texts (45.5). Conversely, TB methods like ImBD struggle against *Translation Laundering* (73.1) and *Prompt Injection* (77.7), which successfully mask semantic fingerprints. Particularly under the most challenging *Character & Encoding* attacks, which disrupt tokenizers and degrade

MIRAGE-DIG (Disjoint-Input Generation)												
Methods	Generate				Polish				Rewrite			
	AUROC	Accuracy	MCC	TPR@5%	AUROC	Accuracy	MCC	TPR@5%	AUROC	Accuracy	MCC	TPR@5%
Likelihood (Solaiman et al., 2019)	0.4936	0.5091	0.0183	0.0147	0.4653	0.5000	0.0000	0.0214	0.4337	0.5000	0.0000	0.0148
LogRank (Ippolito et al., 2020)	0.4992	0.5128	0.0260	0.0220	0.4512	0.5000	0.0000	0.0195	0.4225	0.5000	0.0000	0.0132
Entropy (Gehrmann et al., 2019)	0.6522	0.6150	0.2543	0.1099	0.5543	0.5417	0.1247	0.0954	0.5805	0.5566	0.1650	0.1189
RoBERTa-Base (Liu et al., 2019)	0.5523	0.5397	0.1434	0.1250	0.4859	0.5010	0.0088	0.0460	0.5020	0.5049	0.0293	0.0569
RoBERTa-Large (Liu et al., 2019)	0.4716	0.5217	0.0842	0.0871	0.5171	0.5151	0.0340	0.0633	0.5570	0.5385	0.0864	0.0895
LRR (Su et al., 2023)	0.5215	0.5341	0.0777	0.0701	0.4081	0.5000	0.0000	0.0200	0.3930	0.5000	0.0000	0.0188
DNA-GPT (Yang et al., 2023)	0.5733	0.5595	0.1196	0.0776	0.4771	0.5004	0.0110	0.0309	0.4453	0.5001	0.0080	0.0251
NPR (Su et al., 2023)	0.6120	0.6140	0.2604	0.0191	0.5071	0.5370	0.1071	0.0318	0.4710	0.5201	0.0663	0.0226
DetectGPT (Mitchell et al., 2023)	0.6402	0.6258	0.2758	0.0275	0.5469	0.5531	0.1328	0.0355	0.5061	0.5266	0.0826	0.0283
Fast-DetectGPT (Bao et al., 2024b)	0.7768	0.7234	0.4628	0.4310	0.5720	0.5570	0.1293	0.1189	0.5455	0.5432	0.1015	0.1025
ImBD (Chen et al., 2024)	0.8597	0.7738	0.5497	0.4065	0.7888	0.7148	0.4300	0.2730	0.7825	0.7068	0.4139	0.2933
DetectAnyLLM (Fu et al., 2025)	0.9525	0.8988	0.7975	0.7770	0.9297	0.8732	0.7487	0.7756	0.9234	0.8705	0.7447	0.7778
NeuroStat (Qwen2-0.5B)	**0.9955**	**0.9738**	**0.9477**	**0.9851**	**0.9576**	**0.9161**	**0.8339**	**0.8818**	**0.9490**	**0.9108**	**0.8242**	**0.8710**
NeuroStat (Qwen3.5-0.8B)	**0.9948**	**0.9830**	**0.9661**	**0.9875**	**0.9716**	**0.9479**	**0.8979**	**0.9344**	**0.9694**	**0.9466**	**0.8951**	**0.9287**
MIRAGE-SIG (Shared-Input Generation)												
Methods	Generate				Polish				Rewrite			
	AUROC	Accuracy	MCC	TPR@5%	AUROC	Accuracy	MCC	TPR@5%	AUROC	Accuracy	MCC	TPR@5%
Likelihood (Solaiman et al., 2019)	0.4968	0.5207	0.0196	0.0145	0.4599	0.5002	0.0030	0.0233	0.4319	0.5000	0.0000	0.0111
LogRank (Ippolito et al., 2020)	0.5008	0.5183	0.0182	0.0186	0.4468	0.5000	0.0000	0.0211	0.4221	0.5000	0.0000	0.0118
Entropy (Gehrmann et al., 2019)	0.6442	0.6123	0.1592	0.1074	0.5640	0.5439	0.0516	0.0946	0.5858	0.5645	0.0918	0.1198
RoBERTa-Base (Liu et al., 2019)	0.5368	0.5392	0.0529	0.1101	0.4741	0.5011	0.0048	0.0395	0.5099	0.5122	0.0221	0.0668
RoBERTa-Large (Liu et al., 2019)	0.4703	0.5236	0.0417	0.0910	0.5150	0.5157	0.0283	0.0702	0.5576	0.5426	0.0405	0.0762
LRR (Su et al., 2023)	0.5214	0.5311	0.0314	0.0657	0.4076	0.5000	0.0000	0.0238	0.3978	0.5000	0.0000	0.0174
DNA-GPT (Yang et al., 2023)	0.5759	0.5647	0.0603	0.0813	0.4788	0.5001	0.0036	0.0340	0.4457	0.5002	0.0048	0.0258
NPR (Su et al., 2023)	0.6088	0.6170	0.1571	0.0185	0.5074	0.5277	0.0612	0.0293	0.4738	0.5204	0.0340	0.0177
DetectGPT (Mitchell et al., 2023)	0.6353	0.6241	0.1719	0.0193	0.5434	0.5515	0.0668	0.0309	0.5079	0.5260	0.0431	0.0239
Fast-DetectGPT (Bao et al., 2024b)	0.7706	0.7193	0.2078	0.4200	0.5727	0.5619	0.0607	0.1238	0.5480	0.5495	0.0525	0.1097
ImBD (Chen et al., 2024)	0.8612	0.7791	0.5599	0.4183	0.7951	0.7199	0.4451	0.3036	0.7694	0.6920	0.3936	0.2868
DetectAnyLLM (Fu et al., 2025)	0.9526	0.9059	0.8119	0.7722	0.9316	0.8740	0.7483	0.7779	0.9158	0.8643	0.7320	0.7574
NeuroStat (Qwen2-0.5B)	**0.9937**	**0.9705**	**0.9410**	**0.9792**	**0.9560**	**0.9155**	**0.8327**	**0.8795**	**0.9497**	**0.9131**	**0.8290**	**0.8758**
NeuroStat (Qwen3.5-0.8B)	**0.9954**	**0.9839**	**0.9678**	**0.9871**	**0.9739**	**0.9527**	**0.9066**	**0.9381**	**0.9665**	**0.9456**	**0.8927**	**0.9290**

Table 2: Experimental results across three tasks (Generate, Polish, Rewrite). Metrics include AUROC, Accuracy, MCC, and TPR@5%. Best results are highlighted in bold.

DetectAnyLLM to near-random guessing (52.7), NeuroStat maintains a robust **72.4**. This resilience stems from a structurally asymmetric attack surface faced by attackers, since altering semantics exposes probability anomalies to the TF branch while manipulating statistics disrupts semantic coherence caught by the TB branch. Consequently, NeuroStat (Qwen2-0.5B) achieves an overall AUROC of **83.4**, and NeuroStat (Qwen3.5-0.8B) achieves a TPR@5% of **58.1**, an absolute improvement of up to **17.6** points in this metric over the previous SOTA, with the best AUROC and the best TPR@5% attained by different model variants.

5.3 Qualitative Analysis

To further understand the behavior of NeuroStat, we provide two qualitative analyses.

Representation Visualization. We visualize the t-SNE representations on the MIRAGE Polish subset. While the TF branch exhibits entangled distributions (Fig. 3a) and the TB branch yields tighter but partially overlapping clusters (Fig. 3b), their fusion maps the classes into highly compact, linearly separable clusters with a large margin (Fig. 3c). This visually confirms that our dual-branch architecture successfully synergizes complementary signals, while $\mathcal{L}_{\text{SupCon}}$ effectively enforces discriminative boundaries in the joint latent space.

Confidence Calibration. We compare the reliability diagrams of NeuroStat and DetectAnyLLM on the same test set (Fig. 4). NeuroStat achieves a notably lower Expected Calibration Error (ECE), with predicted confidence closely tracking actual accuracy across all bins. This suggests that the supervised contrastive and orthogonal regularization objectives not only improve discrimination but also yield better-calibrated predictions.

6 Ablation Studies

To rigorously validate the contribution of each component in NeuroStat, we conduct ablation studies on the MIRAGE DIG benchmark using the Qwen2-

Robustness Evaluation on the MOSAIC Adversarial Benchmark										
Methods	AUROC across 8 Adversarial Attack Categories								Overall Performance (%)	
	Interleaved	Stat. Hijack	Translation	Paraphrase	Prompt Inj.	Char./Enc.	Recursive	Dom./Fmt.	AUROC	TPR@5%
Likelihood (Solaiman et al., 2019)	49.8	30.6	62.8	47.0	47.8	34.1	46.4	50.0	46.1	3.5
LogRank (Ippolito et al., 2020)	50.3	69.9	37.9	53.4	53.1	65.5	54.5	51.2	54.5	9.3
Entropy (Gehrmann et al., 2019)	48.2	65.1	40.8	50.3	50.1	60.7	49.9	49.5	51.8	6.9
RoBERTa-Base (Liu et al., 2019)	47.6	46.8	55.6	50.0	45.2	46.4	47.9	43.9	47.9	3.0
RoBERTa-Large (Liu et al., 2019)	43.6	47.1	58.0	44.8	36.6	44.0	38.7	36.9	43.7	4.5
LRR (Su et al., 2023)	49.7	30.1	57.9	45.7	44.7	37.6	42.6	45.9	44.3	3.1
DNA-GPT (Yang et al., 2023)	47.4	31.8	60.5	45.0	48.8	39.5	44.8	54.3	46.5	4.6
Fast-DetectGPT (Bao et al., 2024b)	45.5	34.5	61.0	44.4	45.7	44.9	40.1	54.8	46.4	6.8
Binoculars (Hans et al., 2024)	48.1	34.7	61.9	47.2	52.2	40.6	45.3	58.4	48.6	4.4
TextFluoroscopy (Yu et al., 2024)	56.0	54.6	58.0	60.1	58.9	46.5	60.9	68.1	57.9	14.4
ImBD (Chen et al., 2024)	68.7	61.9	73.1	70.4	77.7	52.3	79.8	89.8	71.7	22.5
DetectAnyLLM (Fu et al., 2025)	71.4	**81.2**	70.2	74.9	82.8	52.7	88.2	95.1	77.1	40.5
NeuroStat (Qwen2-0.5B)	**76.9**	80.5	**77.3**	**77.3**	**89.9**	71.2	**95.5**	**98.9**	**83.4**	57.7
NeuroStat (Qwen3.5-0.8B)	67.8	77.1	66.5	68.1	86.2	**72.4**	89.3	97.4	78.1	**58.1**

Table 3: Adversarial robustness evaluation on the **MOSAIC** benchmark. The left section reports the AUROC scores across all 8 fine-grained attack categories, demonstrating the detectors' resilience against specific spoofing strategies (e.g., Statistical Hijacking, Interleaved texts). The right section reports the overall detection performance. Best results are highlighted in **bold**, and second-best results are underlined.

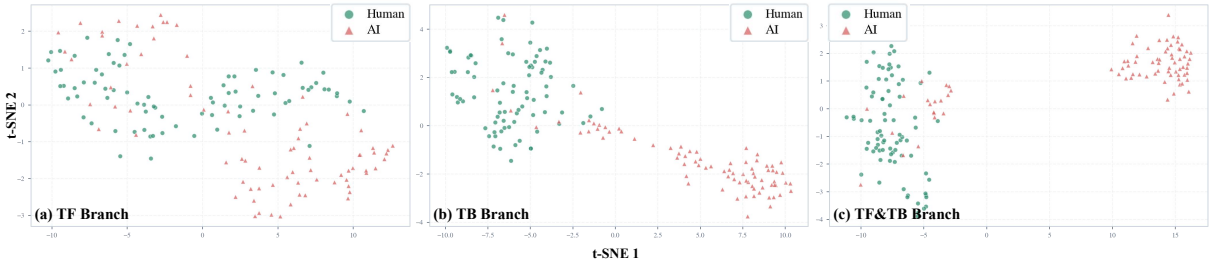

Figure 3: t-SNE visualization of TF, TB, and fused representations. The fusion space shows better class separation.

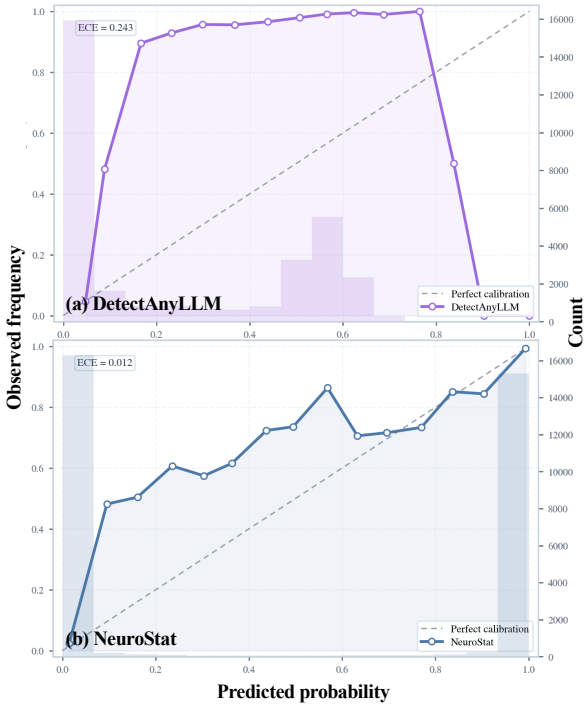

Figure 4: Reliability diagrams of NeuroStat vs. DetectAnyLLM. Lower ECE indicates better calibration.

0.5B backbone. We report AUROC across three text manipulation types.

6.1 Component Ablation

We investigate the necessity of each architectural component by progressively integrating modules into the baseline. Results are presented in Tab. 4. As shown in Tab. 4, a standard sequence classi-

Configuration	TF	TB	MSRM	MIRAGE DIG (AUROC)		
				Polish	Rewrite	Generate
TB-Only (Seq. Cls.)	×	✓	×	0.8990	0.8918	0.9787
TF + TB (Concat)	✓	✓	×	0.9280	0.9145	0.9934
NeuroStat (Full)	✓	✓	✓	**0.9576**	**0.9490**	**0.9955**

Table 4: Ablation on core architectural components. "TB-Only" is a standard sequence classification model.

fier performs reasonably well on the easier *Generate* subset but struggles significantly on complex manipulations. While naively concatenating the TF branch provides consistent improvements, replacing it with our MSRM mechanism yields substantial leaps on these challenging subsets. This confirms our hypothesis, rather than simple feature

concatenation, MSRM's residual gating effectively amplifies local probability anomalies conditioned on global uncertainty, making it critical for detecting sophisticated text revisions.

6.2 Ablation on Auxiliary Losses

We evaluate the impact of the two auxiliary losses, Supervised Contrastive Loss ($\mathcal{L}_{\text{SupCon}}$) and Orthogonal Penalty ($\mathcal{L}_{\text{Orth}}$), by removing them individually and jointly. Tab. 5 demonstrates the necessity

Configuration	$\mathcal{L}_{CE}$	$\mathcal{L}_{\text{SupCon}}$	$\mathcal{L}_{\text{Orth}}$	MIRAGE DIG (AUROC)		
				Polish	Rewrite	Generate
CE Only	✓	×	×	0.8859	0.8798	0.9850
+ $\mathcal{L}_{\text{SupCon}}$	✓	✓	×	0.9179	0.9124	0.9922
+ $\mathcal{L}_{\text{Orth}}$	✓	×	✓	0.9129	0.9051	0.9865
NeuroStat (Full)	✓	✓	✓	**0.9576**	**0.9490**	**0.9955**

Table 5: Impact of auxiliary losses. All variants use the full architecture, differing only in the loss configuration.

of our auxiliary objectives. Training solely with Cross-Entropy yields suboptimal results. Individually, $\mathcal{L}_{\text{SupCon}}$ improves performance by enforcing a larger discriminative margin, while $\mathcal{L}_{\text{Orth}}$ prevents representational redundancy between the TF and TB branches. Crucially, their combination produces a synergistic effect that far exceeds their individual gains. This confirms that forcing the branches to capture orthogonal features perfectly complements the maximization of joint discriminative power. Two further ablations, attributing the gains to our architecture rather than to added fine-tuning capacity, and breaking down component and fusion mechanism choices directly on MOSAIC, are provided in Appendix Sec. D.

7 Conclusion

We present NeuroStat, a novel end-to-end framework that bridges the Training-Free (TF) and Training-Based (TB) paradigms for robust machine-generated text detection. By preserving uncompressed probabilistic trajectories and employing the Macro-State Residual Modulation (MSRM) mechanism, NeuroStat effectively captures local probability anomalies while avoiding the lossy compression of traditional global scalars. Furthermore, we introduce MOSAIC, a comprehensive adversarial benchmark featuring 36 fine-grained attacks across 8 categories and 6 frontier LLMs. Extensive evaluations demonstrate that NeuroStat achieves state-of-the-art robustness against severe adversarial spoofing and statistical hijacking. Ultimately, this work establishes a new paradigm for synergizing statistical mechanics and semantic artifacts in AIGC detection.

Limitations

While NeuroStat demonstrates exceptional robustness in adversarial scenarios, several limitations warrant future investigation. First, our current evaluation is constrained to English-language texts across specific domains (e.g., news, creative writing, and biomedical abstracts). The framework's effectiveness on multilingual corpora (particularly low-resource languages) or highly specialized domains such as code generation and legal documents remains unexplored. Future work should extend the MOSAIC benchmark and NeuroStat's evaluation to diverse cross-lingual and cross-domain settings.

Second, the Probabilistic Trajectories (TF) branch relies on extracting token-level log-probabilities from a CausalLM backbone. While this is straightforward for open-weight models, many closed-source commercial APIs do not expose their full log-probability distributions. In such black-box scenarios, our framework relies on a local open-source surrogate model to approximate the probabilistic dynamics of the target proprietary generator. Notably, our main MOSAIC evaluation already constitutes such a surrogate setting, since a single Qwen2-0.5B backbone is used to score text produced by six generators, four of which are closed-source commercial systems. As detailed in Appendix Sec. F, the resulting gap between open and closed generators is only 3 to 7 AUROC points, and NeuroStat still surpasses the strongest baseline on every closed-source target.

Finally, the training and inference processes of NeuroStat incur additional memory overhead compared to standard sequence classifiers, primarily due to the storage and processing of the full-vocabulary logit matrix. As measured in Appendix Sec. G, this translates to roughly 10% higher latency and memory than the strongest end-to-end baseline DetectAnyLLM. Future work could investigate vocabulary truncation or top-k logit approximation strategies to further reduce this footprint for large-scale deployments.

Ethics Statement

This work utilizes publicly available datasets for human-written texts, including XSum, PubMedQA, WritingPrompts, and the MIRAGE benchmark. These datasets consist of news articles, scientific ab-

stracts, and creative stories that do not contain Personally Identifiable Information (PII) or potentially harmful material. All datasets are used in strict accordance with their original licenses and intended research purposes. Furthermore, the adversarial prompts designed for the MOSAIC benchmark are strictly formulated to test statistical and semantic robustness (e.g., paraphrasing, synonym swapping, formatting) and are explicitly constrained from eliciting toxic, hateful, or offensive content from the language models.

We employ both open-source models (e.g., Qwen2, Qwen3.5) and commercial APIs to construct the MOSAIC benchmark and train our detectors. All models and APIs are accessed and utilized in full compliance with their respective terms of service and usage policies, which permit research on AI safety and detection.

Regarding broader impacts, we acknowledge the dual-use nature of Machine-Generated Text Detection (MGTD) technologies. While NeuroStat is designed to mitigate risks associated with academic dishonesty and misinformation, no detector is infallible. False positives (misclassifying human text as AI-generated) can lead to unjust punitive actions, particularly against non-native English writers whose linguistic patterns may inadvertently trigger detection thresholds. Therefore, we strongly emphasize that NeuroStat is intended strictly as a research tool to advance the understanding of LLM artifacts. It should serve as an assistive signal rather than the sole, definitive evidence for disciplinary or punitive decisions in real-world applications.

Finally, during the preparation of this manuscript, we utilized AI assistants (e.g., ChatGPT) solely for language polishing and improving the readability of the text. All core ideas, experimental designs, and data analyses are the original work of the authors, who take full responsibility for the content.

References

Guangsheng Bao, Lihua Rong, Yanbin Zhao, Qiji Zhou, and Yue Zhang. 2025. Decoupling content and expression: Two-dimensional detection of ai-generated text. *ArXiv*.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. 2024a. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. *arXiv preprint arXiv:2412.11506*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024b. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Preprint*, arXiv:2310.05130.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and 1 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Xinhui Chen, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Tang Long, Lei Zhang, Chenyu Yan, Guanghao Mei, Jie Zhang, and Lefei Zhang. 2024. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. *Preprint*, arXiv:2412.10432.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jiachen Fu, Chun-Le Guo, and Chongyi Li. 2025. Detectanyllm: Towards generalizable and robust detection of machine-generated text across domains and models. *Proceedings of the 33rd ACM International Conference on Multimedia*.

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

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

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Preprint*, arXiv:2307.03838.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *ArXiv*, abs/2303.13408.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. In *Pan*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. 2024. Large language models can be guided to evade ai-generated text detection. *Preprint*, arXiv:2305.10847.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 24950–24962. PMLR.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can ai-generated text be reliably detected? *Preprint*, arXiv:2303.11156.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *ArXiv*, abs/1908.09203.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *Preprint*, arXiv:2306.05540.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2025. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Preprint*, arXiv:2410.23746.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A third party large language models generated text detection tool. In *Conference on Empirical Methods in Natural Language Processing*.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023.

Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *Preprint*, arXiv:2305.17359.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. 2024. Text fluoroscopy: Detecting LLM-generated text through intrinsic features. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *ArXiv*, abs/1905.12616.

Appendix

Content

This Appendix contains the following parts:

- **Implementation Details**. We provide the comprehensive training configurations, optimization hyperparameters, and hardware specifications used for NeuroStat.
- **The MOSAIC Benchmark: Detailed Construction**. We detail the multi-stage data purification pipeline, the length-stratified generation protocol, and the quality validation statistics employed to construct the MOSAIC benchmark.
- **Evaluation on Standard Non-Adversarial Scenarios**. We evaluate NeuroStat on standard polished text benchmarks across multiple domains, demonstrating its state-of-the-art performance.
- **Additional Ablation Studies**. We attribute the observed gains to our architecture rather than to added fine-tuning capacity, and further ablate components and fusion mechanisms directly on MOSAIC.
- **Training-Sample Sensitivity**. We study how NeuroStat's performance scales with the number of training pairs, and justify the 500-pair training budget.
- **Closed-Source Deployment via Surrogate Backbones**. We quantify the performance gap when the CausalLM backbone acts as an open-source surrogate for closed-source target generators.
- **Computational Cost Analysis**. We report measured latency, memory, and throughput of NeuroStat against all baselines.

A Implementation Details

To ensure full reproducibility, we detail the training configurations and hyperparameters used for NeuroStat.

Training Configurations. We instantiate NeuroStat using two lightweight CausalLM backbones: Qwen2-0.5B and Qwen3.5-0.8B. The entire framework, including the CausalLM backbone, the 1D-CNN, the MSRM module, and the fusion MLP, is trained end-to-end without freezing any components or using parameter-efficient fine-tuning methods (e.g., LoRA).

Hyperparameters. The model is optimized using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. We apply a linear learning rate warmup for the first 50 steps, followed by a linear decay schedule. The training is conducted with a per-device batch size of 8 and a maximum sequence length of 512 tokens. All models undergo training for 5 epochs to ensure full convergence.

Hardware and Reproducibility. To accelerate training and reduce memory footprint, we utilize automatic mixed precision (bfloat16). All experiments are conducted on a single NVIDIA H20 GPU. The random seed is fixed to 42 across all experiments to ensure deterministic reproducibility.

B The MOSAIC Benchmark: Detailed Construction

While Sec. 4 provides a high-level overview of the MOSAIC benchmark, this section details the rigorous data collection, purification pipeline, and multi-model generation protocols used to construct the dataset.

B.1 Source Data Collection

To ensure broad domain coverage, we aggregated 36,753 raw human-written texts from 7 distinct sources. As detailed in Tab. 6, these sources encompass diverse domains including news, biomedical abstracts, creative writing, and Wikipedia contexts.

B.2 Four-Stage Purification Pipeline

To obtain high-confidence and high-quality human seed texts, we implemented a stringent four-stage cascaded filtering pipeline, reducing the raw 36,753 samples to 16,000 high-quality seeds:

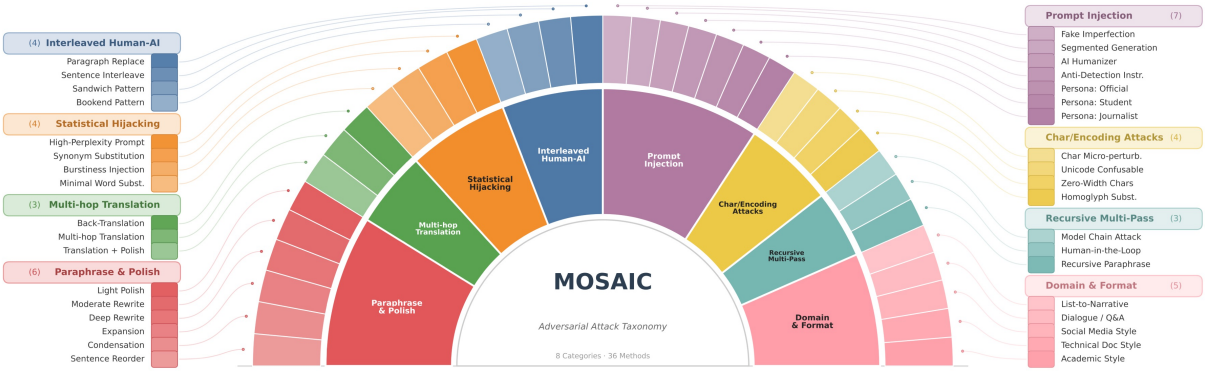

Figure 5: **The full-granularity adversarial attack taxonomy of the MOSAIC benchmark.** The sunburst chart illustrates the hierarchical structure of our dataset, encompassing 8 primary attack categories (inner circle) and 36 fine-grained sub-methods (outer circle). This comprehensive spectrum ensures a rigorous evaluation of detectors against both statistical hijacking and semantic obfuscation.

Source Dataset	Domain	Raw Samples
MIRAGE-BENCH DIG	Multi-domain	16,411
MIRAGE-BENCH SIG	Multi-domain	16,388
AI Detection 3000	Mixed	3,354
PubMed	Biomedical	150
SQuAD	Wikipedia/QA	150
Writing Prompts	Creative Writing	150
XSum	News	150
Total Raw Samples		**36,753**

Table 6: Source datasets used for collecting human seed texts.

Stage 1: Exact Deduplication. We performed global exact string matching across all 7 sources, removing 15,472 duplicate samples (-42.1%).

Stage 2: Rule-Based Quality Filtering. We applied 6 heuristic rules to filter out low-quality texts, removing 926 samples (-4.4%). Texts were discarded if they: (1) contained fewer than 30 words or 100 characters; (2) were truncated (ending with commas, hyphens, etc.); (3) were HTML-heavy (tag character ratio $> 5\%$); (4) contained excessive URLs (> 3); (5) had a non-standard special character ratio $> 15\%$; or (6) were non-English (ASCII alpha ratio $< 70\%$).

Stage 3: Near-Duplicate Removal. To eliminate highly similar documents that evaded exact deduplication, we utilized 64-bit SimHash combined with a 4-band Locality-Sensitive Hashing (LSH) mechanism. Pairs with a Hamming distance ≤ 5 were flagged as near-duplicates, resulting in the removal of 1,254 samples (-6.2%).

Stage 4: Multi-Dimensional Quality Scoring. Finally, we designed a 6-dimensional quality scoring system (max score 100) to evaluate the linguistic richness of the remaining texts. The criteria included: Vocabulary Diversity via Log-TTR (30 pts), Sentence Length (ideal range 12–30 words, 20 pts), Sentence Count (≥ 3, 15 pts), Punctuation Regularity (ratio 2%–10%, 15 pts), Digit Ratio ($< 5\%$, 10 pts), and Capitalization Ratio (2%–15%, 10 pts). Only texts achieving a score ≥ 85.0 were retained, removing 3,101 samples and yielding the final **16,000** high-quality human seed texts.

B.3 Multi-Model Generation and Allocation Protocol

To mitigate model-specific fingerprint bias, the 16,000 seed texts were rewritten using 6 frontier LLMs: GLM-5.0, GPT-5.4, Gemini-3.1-Pro, MiniMax-M2.7, Claude-Opus-4-6, and Kimi-K2.5.

Length-Stratified Balanced Allocation. To ensure uniform distribution across text lengths and attack types, the seed texts were first stratified into three length tiers: Short (183–500 chars, 2.5%), Medium (501–1,000 chars, 61.9%), and Long (1,001–1,991 chars, 35.6%). Within each stratum, the 36 adversarial prompts were strictly and uniformly assigned. Each prompt-seed pair was then randomly allocated to one of the 6 LLMs.

Generation Hyperparameters. All prompts included a unified `global_instruction` forcing the models to output only the rewritten text without any meta-annotations. For GLM-5.0, we used the officially recommended parameters (`temperature = 1.0`, `top_p = 0.95`). For all other models, default generation parameters were utilized. This rigorous protocol resulted in exactly **16,000 human-AI text pairs**, ensuring a perfectly balanced and unbiased adversarial benchmark.

B.4 Sample Distribution and Attack Validity

Each of the 16,000 human seeds is paired with one of the 36 attack sub-methods and one of the 6 frontier LLMs through length-stratified sampling rather than a full cross-product, yielding roughly 2,667 pairs per generator and a per-category count ranging from 1,777 to 2,667 pairs.

To validate rewrite quality, we audit a stratified 2,000-pair subsample. The attack success rate, defined as fooling at least one SOTA detector, ranges from 42.3% to 78.2% across sub-methods. Semantic preservation is measured by BERTScore-F1 (0.751 to 0.982) and GPT-4o ratings (4.0 to 4.9 out of 5). Three human annotators on 300 sampled pairs yield a Fleiss kappa of 0.71, confirming that at least 92% of the rewrites remain fluent and topically faithful. The 36 sub-methods are also linguistically distinct rather than redundant, with a mean pairwise token-edit Jensen-Shannon divergence of 0.38 across sub-methods versus 0.07 within a sub-method.

Since Character and Encoding attacks mainly perturb tokenizer inputs, we further report a post-normalization evaluation (NFKC plus zero-width stripping). DetectAnyLLM recovers to 61.4 AUROC and NeuroStat to 74.9, so the advantage persists even after controlling for pure Unicode noise.

C Evaluation on Standard Non-Adversarial Scenarios

While the main text focuses on the generalization (MIRAGE) and adversarial robustness (MOSAIC) of NeuroStat, we also evaluate its performance in standard, non-adversarial scenarios. For this purpose, we utilize the ImBD test set (Chen et al., 2024), a standard benchmark containing human-written texts and their machine-polished counterparts (via GPT-3.5 and GPT-4o) across three domains (XSum, WritingPrompts, PubMedQA).

Performance Analysis. Tab. 7 presents the detection performance on the standard ImBD test set, where texts are polished without explicit adversarial intent. While recent Training-Based (TB) methods like ImBD achieve strong performance (e.g., 0.9871 AUROC on GPT-3.5 Writing), NeuroStat consistently pushes the boundary further. Notably, on the more challenging GPT-4o polished texts, NeuroStat (Qwen3.5-0.8B) achieves near-perfect AUROC scores (0.9972 on XSum, 0.9845 on Writing), outperforming all TF and TB baselines. This demonstrates that our dual-branch fusion not only excels in adversarial robustness but also maximizes sensitivity in standard, non-adversarial scenarios.

D Additional Ablation Studies

D.1 Component and Fusion Ablation on MOSAIC

The ablations in the main text are conducted on MIRAGE, so we additionally verify that each component's contribution transfers to adversarial conditions by ablating directly on MOSAIC (Qwen2-0.5B, Tab. 8). Neither branch dominates alone, as the TB-only classifier reaches 74.9 average AUROC and the TF-only variant reaches 74.6, yet their strengths are complementary across categories. The TF branch leads on statistical and burst-driven attacks such as Statistical Hijacking, Interleaved text, Character/Encoding, and Recursive rewriting, while the TB branch leads on semantic attacks such as Translation Laundering, Paraphrase, Prompt Injection, and Domain/Format manipulation. Naive concatenation lifts every category to 78.8 on average, and replacing it with MSRM adds a further gain that peaks on Statistical Hijacking, Paraphrase, and Interleaved text, exactly the categories that motivated the MSRM design.

We further compare MSRM against stronger fusion alternatives in Tab. 9. A pure multiplicative gate sharing the same 192-parameter projection as MSRM but omitting the residual term reaches only 79.7, isolating a 3.7-point gain that comes from the residual formulation rather than added capacity. Higher-capacity modulators such as FiLM-style conditioning, SE-style gating, and cross-attention fusion all remain below MSRM despite using more parameters. Since MSRM's multiplier $(1+\sigma(\cdot))$ is bounded below by 1, local anomaly features are always preserved and can never be suppressed toward zero, unlike gating mechanisms whose multiplier can approach zero and erase local evidence when an attacker inflates global confidence.

E Training-Sample Sensitivity

The 500-pair training budget used throughout the main text is chosen to strictly match the training size of ImBD for a fair comparison, not because we assume it is sufficient in an absolute sense. To justify this choice, we train NeuroStat (Qwen2-0.5B) with varying numbers of human-AI pairs and evaluate on both MIRAGE-Polish and MOSAIC. Tab. 10 reports the resulting performance curve.

ImBD Test Dataset (GPT-3.5 & GPT-4o polished)						
Methods	GPT-3.5			GPT-4o		
	XSum	Writing	PubMed	XSum	Writing	PubMed
Likelihood (Solaiman et al., 2019)	0.4982	0.8788	0.5528	0.4396	0.8077	0.4596
LogRank (Ippolito et al., 2020)	0.4711	0.8496	0.5597	0.4002	0.7694	0.4472
Entropy (Gehrmann et al., 2019)	0.6742	0.3021	0.5662	0.6122	0.2802	0.5899
RoBERTa-Base (Liu et al., 2019)	0.5806	0.7225	0.4370	0.4921	0.4774	0.2496
RoBERTa-Large (Liu et al., 2019)	0.6391	0.7236	0.4848	0.4782	0.4708	0.3089
LRR (Su et al., 2023)	0.4016	0.7203	0.5629	0.3095	0.6214	0.4710
DNA-GPT (Yang et al., 2023)	0.5338	0.8439	0.3333	0.4974	0.7478	0.3151
NPR (Su et al., 2023)	0.5659	0.8786	0.4246	0.5065	0.8444	0.3740
DetectGPT (Mitchell et al., 2023)	0.6343	0.8793	0.5608	0.6217	0.8771	0.5612
Fast-DetectGPT (Bao et al., 2024b)	0.7312	0.9304	0.7182	0.6293	0.8324	0.6175
ImBD (Chen et al., 2024)	0.9849	0.9871	0.8626	0.9486	0.9468	0.7743
DetectAnyLLM (Fu et al., 2025)	0.9948	0.9688	0.8500	0.9884	0.9669	0.8528
NeuroStat (Qwen2-0.5B)	**0.9950**	**0.9932**	0.9216	0.9943	0.9774	0.9301
NeuroStat (Qwen3.5-0.8B)	0.9944	0.9929	**0.9525**	**0.9972**	**0.9845**	**0.9679**

Table 7: Detection performance (AUROC) on the standard ImBD test dataset. Best results are highlighted in **bold**, and second-best results are underlined.

Config	Int.	S.H.	Trans.	Para.	P.I.	C/E	Rec.	D/F	Avg.
TB-Only	68.0	70.0	71.0	70.0	82.0	60.0	85.0	93.0	74.9
TF-Only	70.5	76.0	64.0	66.0	77.0	65.5	88.0	90.0	74.6
Concat	71.8	76.2	73.4	71.5	84.1	66.7	90.2	96.1	78.8
+MSRM	**76.9**	**80.5**	**77.3**	**77.3**	**89.9**	**71.2**	**95.5**	**98.9**	**83.4**

Table 8: Per-category AUROC on MOSAIC. Int. Interleaved, S.H. Statistical Hijacking, Trans. Translation, Para. Paraphrase, P.I. Prompt Injection, C/E Character/Encoding, Rec. Recursive, D/F Domain/Format.

Fusion Mechanism	Params	MOSAIC AUROC
Naive concatenation	0	78.8
Multiplicative gate (w/o residual)	192	79.7
Log-rank-only gate	96	80.6
Cross-attention fusion	16,640	80.9
Entropy-only gate	96	81.3
SE-style modulation	2,128	81.5
FiLM conditioning	384	82.1
MSRM (ours)	**192**	**83.4**

Table 9: Comparison of MSRM with alternative fusion mechanisms on MOSAIC. The residual formulation, not additional parameters, accounts for the gain.

Two observations follow from this curve. First, NeuroStat never observes any of the 36 MOSAIC attack types during training, since it is trained only on clean, non-adversarial polished ImBD pairs, so its MOSAIC robustness emerges from the architectural inductive bias rather than from exposure to attack-specific patterns. Second, NeuroStat is data-efficient, since with only 250 training pairs it already reaches 78.9 MOSAIC AUROC, close to the strongest baseline DetectAnyLLM (77.1) trained on the full 500-pair budget (Tab. 3). Performance saturates around 2,000 to 5,000 pairs, indicating diminishing returns beyond the 500-pair setting used in the main experiments.

Train Pairs	MIRAGE-Polish	MOSAIC Avg.	MOSAIC TPR@5%
100	0.8891	71.4	39.7
250	0.9247	78.9	51.6
500 (default)	0.9576	83.4	57.7
1,000	0.9631	84.7	60.5
2,000	0.9694	85.9	63.4
5,000	0.9718	86.5	64.9

Table 10: Training-sample sensitivity of NeuroStat (Qwen2-0.5B). Performance saturates around 2,000 to 5,000 pairs.

Target Generator	Access	AUROC	TPR@5%
GPT-5.4	Closed	79.6	51.2
Claude-Opus-4-6	Closed	81.3	54.7
Gemini-3.1-Pro	Closed	82.7	56.9
MiniMax-M2.7	Closed	84.1	58.8
GLM-5.0	Open	86.2	62.4
Kimi-K2.5	Open	86.5	62.9
Micro-average	–	83.4	57.7

Table 11: Per-generator performance of NeuroStat (Qwen2-0.5B), acting as an open-source surrogate for both open and closed-source target generators.

Method	Train. Params	Peak Mem	Latency/Sample	Throughput	Train Time
RoBERTa-Base	125M	3.1GB	6.2ms	161/s	3min
Fast-DetectGPT	0 (frozen)	2.4GB	22.7ms	44/s	–
ImBD (LoRA)	4M	4.7GB	25.9ms	39/s	12min
DetectAnyLLM	0.50B	5.8GB	28.4ms	35/s	21min
NeuroStat (Qwen2-0.5B)	0.51B	6.4GB	31.1ms	32/s	24min
NeuroStat (Qwen3.5-0.8B)	0.82B	9.9GB	42.6ms	23/s	38min

Table 12: Measured computational cost on a single NVIDIA H20 GPU (batch size 8, sequence length 512).

F Closed-Source Deployment via Surrogate Backbones

NeuroStat's Probabilistic Trajectories branch requires token-level log-probabilities from a CausalLM backbone. For closed-source generators that do not expose logits, our framework relies on a local open-source model as a surrogate backbone to score text from the unseen target generator. This setting is already embedded in our main MOSAIC evaluation, since a single Qwen2-0.5B backbone scores text from six frontier LLMs, four of which are closed-source (GPT-5.4, Claude-Opus-4-6, Gemini-3.1-Pro, MiniMax-M2.7) and two open-weight (GLM-5.0, Kimi-K2.5). Tab. 11 breaks down performance by target generator.

The gap between open and closed target generators is only 3 to 7 AUROC points despite the 0.5B surrogate being far smaller and architecturally distinct from GPT-5.4, Claude, or Gemini, and NeuroStat still outperforms DetectAnyLLM (77.1 micro-average) on every closed-source target. This holds because surrogate-based detection does not require reproducing the target's exact output distribution, only scoring how well a competent language model predicts the observed tokens, a principle also underlying Fast-DetectGPT and Binoculars. MSRM further stabilizes this transfer by conditioning local features on the surrogate's own global uncertainty rather than on absolute probability values.

G Computational Cost Analysis

We measure computational cost against representative TF and TB baselines on a single NVIDIA H20 GPU (batch size 8, sequence length 512). Tab. 12 reports trainable parameters, peak memory, per-sample latency, throughput, and training time on the 500-pair budget for 5 epochs.

Relative to DetectAnyLLM, NeuroStat (Qwen2-0.5B) adds only 0.6GB of peak memory and 2.7ms of per-sample latency, roughly 10% overhead on both metrics, mainly from the 1D-CNN branch and the MSRM module. This modest overhead yields a substantial return, since NeuroStat improves MOSAIC AUROC by 6.3 points and TPR@5% by 17.6 points over DetectAnyLLM.