

PAWN++ in the Wild: Evaluating AI-Generated Text Detection on the Web

Oleg Chernikov Evgenii Ptitsyn Albina Burlova

HSE University

{ovchernikov, egptitsyn}@edu.hse.ru

aburlova@hse.ru

Abstract

Machine-generated text is becoming a routine part of the web. Most detectors, however, are still evaluated on curated datasets rather than real web content. We introduce PAWN++, a compact supervised detector that learns which likelihood signals from two frozen language models are most informative. On MAGE, our method reaches 98.36 AUROC in distribution and 98.42 under a joint domain and generator shift. It also transfers well to external benchmarks without target-specific training.

To test whether this robustness extends to the web, we construct WATTLE, a benchmark of 41,300 texts in which each machine output is tied to its original human web page. The resulting data capture realistic transformations of web content as well as variation across languages and time. We use WATTLE to compare our method with seven other detectors. Performance deteriorates sharply on realistic rewrites: our method reaches 70.1 AUROC overall and 58.1 on text produced by a frontier model. Likelihood-based detectors fall below chance, and no evaluated system identifies more than 12% of frontier rewrites at a 1% false-positive rate. These results expose a substantial gap between standard evaluation scores and reliable detection on the web.¹

1 Introduction

Large language models write fluent news, reviews, answers, and essays. Their output now flows into the same channels as human writing, and the ability to tell the two apart underpins moderation, provenance, and the curation of training corpora. Detection research has produced strong supervised classifiers and elegant zero shot statistics, and it evaluates them almost exclusively on curated benchmarks assembled from clean sources such as news

¹All data and code are available at <https://github.com/OlegHSE/PAWN-in-the-Wild>

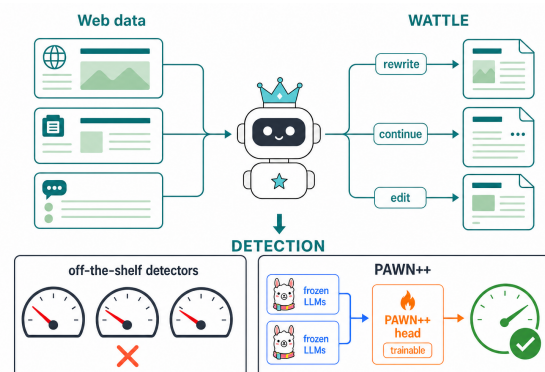


Figure 1: WATTLE anchors machine rewrites, continuations, and edits to real web pages. Likelihood ratio detectors fall below chance on frontier text, and the learned head of PAWN++ reads two frozen probes to stay above chance where fixed statistics invert.

wires, Wikipedia, and question answering forums (Li et al., 2024; Dugan et al., 2024; Artemova et al., 2025).

This paper follows one detector along the full path from a curated benchmark to the open web. We first present PAWN++, an extension of the Perplexity Attention Weighted Network (Miralles-González et al., 2026) that reads token level likelihood evidence from two related frozen language models through a learned head of 2.7M trainable parameters, and it delivers the highest average recall in its MAGE comparison (Li et al., 2024) with competitive in distribution AUROC, and the best transfer to unseen domains and generators. Applied without any adaptation, the same checkpoint leads the no training comparison on RAID (Dugan et al., 2024) at a fixed false positive budget, attains the strongest AUROC in its HC3 comparison in both English and Chinese (Guo et al., 2023), a language it never saw in training, and stays competitive on Beemo (Artemova et al., 2025).

The last and hardest step is the web itself. Web pages are noisy, multilingual, and unlabeled, and

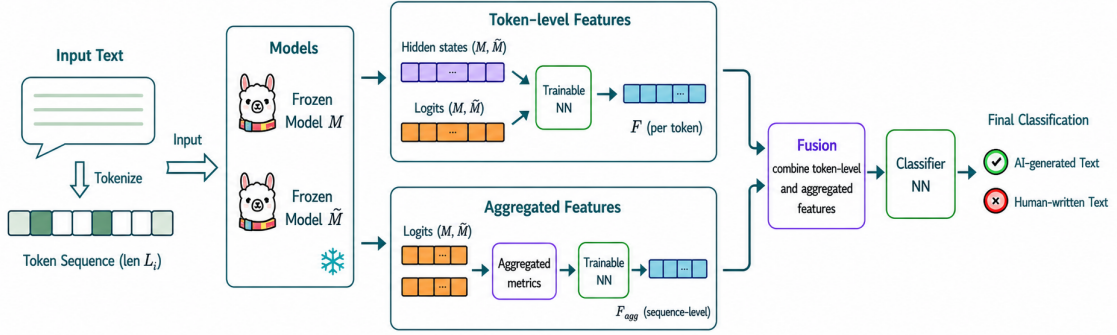


Figure 2: PAWN++ architecture. Two frozen LLMs contribute token probabilities and hidden states; learned PAWN modules combine token, cross-model, and sequence-level evidence.

machine text on the web comes from frontier commercial models. We construct WATTLE, a parallel benchmark of 41,300 objects in which every machine text is anchored to a real web page whose human version predates the modern generation era, covering rewriting, continuation, editing, six languages, and two temporal crawl slices. The results reorder the field. Frozen likelihood ratio and curvature detectors fall below chance on frontier rewrites, supervised detectors keep useful rankings yet collapse at deployment thresholds, and calibrated false positive budgets fail to survive language transfer for every frozen method. A final experiment trains the head anew on five thousand web texts from the benchmark pipeline and shows how misleading in protocol training can be. Scores jump to 96.4 AUROC on rewriting, and the same head trails the frozen checkpoint by 10 to 26 AUROC points on external machine text, so the gains reflect the generation protocol rather than detection.

Our contributions are:

- **PAWN++**, a compact supervised detector that learns to combine token- and sequence-level evidence from two frozen language models. It reaches 98.36 AUROC on MAGE and 98.42 under a joint domain and generator shift, while transferring well to external benchmarks.
- **WATTLE**, a benchmark of 41,300 texts in which every machine output is tied to its original human web page. It captures realistic transformations together with variation across languages and time.

- **A controlled web evaluation** of our method and seven representative baselines. Performance drops sharply on frontier-generated rewrites, where likelihood-based detectors fall below chance and no evaluated method recalls more than 12% at a 1% false-positive rate.

2 Related Work

Supervised detection fine tunes a pretrained encoder into a binary classifier (Liu et al., 2019; Tufts et al., 2025). These models reach high in domain accuracy and lose it under domain, generator, and paraphrase shift (Murgia et al., 2026; Pu et al., 2023; Akhilesh et al., 2025), and RADAR hardens the classifier adversarially against paraphrasing (Hu et al., 2023).

Zero shot detection reads statistics of model likelihoods. DetectGPT uses the curvature of the log probability surface (Mitchell et al., 2023), Fast-DetectGPT replaces its perturbations with an efficient sampling discrepancy (Bao et al., 2024), and Binoculars scores a text by the ratio of perplexity to cross perplexity between two related models (Hans et al., 2024). Follow up work adds token level hidden state discrepancy (Zhu et al., 2026) and temporal or spectral statistics of the surprisal sequence (Basani and Chen, 2026; Luo et al., 2026). The Perplexity Attention Weighted Network sits between the two families, learning how much each token contributes to likelihood features of one frozen model (Miralles-González et al., 2026). PAWN++ extends it with a second frozen probe, cross model metrics, uniform hidden state fusion, and sequence level conditioning (Perez et al., 2018), uniting the

cross model contrast of Binoculars with the surprisal dynamics view of DivEye inside one learned head.

Evaluation resources shaped this field. MAGE spans diverse sources and generators with out of distribution testbeds (Li et al., 2024), RAID stress tests detectors under adversarial attacks at fixed false positive budgets (Dugan et al., 2024), HC3 pairs human and ChatGPT answers in two languages (Guo et al., 2023), and Beemo studies machine text refined by experts and by stronger models (Artemova et al., 2025). A large scale audit of deployed detectors on social media documents reports unstable false positive behavior in the wild (Dutta et al., 2026). Closer to the web itself, WETBench measures task specific machine text on Wikipedia across three languages (Quaremba et al., 2025), DeGenTWeb identifies LLM dominant websites in Common Crawl and search results (He et al., 2026), temporal studies estimate the growth of machine content on the live internet (Dolezal et al., 2026), and detector based filtering of training corpora prevents language model collapse under recursive training (Drayson et al., 2025). WATTLE differs from these resources by anchoring every machine text to a real pre LLM web page, which yields source matched pairs, verified human gold, and an evaluation that reflects the web distribution rather than curated prose.

3 Method

3.1 Original PAWN

PAWN uses next-token distributions from a pre-trained language model, but does not average their statistics uniformly. It learns how much each token should contribute to each likelihood-derived feature (Miralles-González et al., 2026).

LLM pass. A frozen decoder-only LLM processes input $x_{1:L}$ once, returning final hidden states and next-token logits. These outputs can be cached; only the lightweight PAWN networks are optimized.

Metric features. From the next-token distribution p_t , PAWN extracts entropy, maximum log-probability, observed-token log-probability, normalized rank, and cumulative top- p mass:

$$m_t = [\mathcal{H}_t, \max \log p_t, \log p_t(x_{t+1}), r_t, q_t], \quad (1)$$

Here $\mathcal{H}_t$ captures uncertainty, r_t is the observed token’s rank, and q_t is the smallest top- p threshold

Configuration	AUROC
PAWN, Instruct	0.8753
PAWN, Base	0.8953
Two models + metrics	0.9126
Two models + metrics + XPPL	0.9096
Two models + metrics + XPPL + HS	0.9045
Two models + metrics + XPPL + uniform HS	0.9288
PAWN++	0.9309

Table 1: Ablations trained on 5,000 balanced examples and evaluated on the full MAGE test set.

that retains it. A metrics MLP maps m_t to z_t .

Token weighting. At position t , a second MLP uses the current and next token hidden states plus relative position to produce the feature-wise importance-score vector a_t . A softmax over the sequence converts a_t into normalized relative weights, assigning each token a relative importance. The weights are then applied to metric features.

Aggregation and decision. PAWN pools the weighted metric features,

$$F = \sum_{t=1}^{L-1} \text{softmax}_t(a_t) \odot z_t, \quad (2)$$

and a final MLP maps F to a binary logit. PAWN++ retains this metric-weight-aggregate pathway, then adds a second frozen model and sequence-level conditioning.

3.2 Second-Model and Cross-Model Features

Choice of the model pair. Following Binoculars (Hans et al., 2024), PAWN++ uses the Base and Instruct variants of the same model to capture complementary views of a given text. Their shared architecture and tokenizer make token-level likelihood and hidden-state comparisons directly comparable, so score differences reflect the text rather than model-family differences. The pair also provides a cross-model disagreement signal that a single model cannot produce. We do not claim that same-family pairs are optimal, heterogeneous pairs remain for future study.

PAWN++ runs Llama-3.2-1B-Instruct and Llama-3.2-1B in parallel with shared tokenization and frozen weights (Grattafiori et al., 2024). The same five token metrics are computed for each model and concatenated. We also include a cross-perplexity-style feature: the expectation of one model’s negative log-probabilities under the other model’s predictive distribution. Unlike standalone Binoculars, this cross-score is a token feature consumed by the learned PAWN metrics network.

Method	AUROC	Human R.	AI R.	Avg. R.
FastText	83.00	86.34	71.26	78.80
GLTR	74.00	12.42	98.42	55.42
Longformer (fine-tuned)	99.00	82.80	98.27	90.53
DetectGPT*	57.00	86.92	34.05	60.48
RADAR*‡	61.00	52.08	68.56	60.32
RADAR-FT	98.00	85.36	97.26	91.31
Binoculars*‡	65.00	10.60	8.23	9.42
LogRank	68.00	84.91	44.47	64.69
Entropy	67.00	75.68	50.04	62.86
DetectLLM	75.00	64.74	69.02	66.88
FastDetectGPT	70.00	93.65	41.73	67.69
BiScope	86.00	91.54	58.70	75.12
DivEye	88.00	73.72	82.57	78.15
PAWN (GPT2)	98.00	93.75	91.97	92.86
PAWN (Llama3-1B)	99.00	90.68	95.84	93.26
PAWN++ (ours)	98.36	97.20	93.11	95.16

Table 2: MAGE Testbed 4 (all source models and domains). Original baselines are from [Miralles-González et al. \(2026\)](#); additional methods are from [Basani and Chen \(2026\)](#); PAWN++ is our result. * unsupervised; ‡ threshold not selected.

This design represents cross-model disagreement together with each model’s confidence structure rather than treating cross-perplexity as a sufficient statistic.

3.3 Uniform Hidden-State Fusion

The original PAWN weights network uses the last hidden layer. We instead normalize and average all K selected layers for model j :

$$\tilde{h}_t^{(j)} = \frac{1}{K} \sum_{k=1}^K \text{LayerNorm}(h_t^{(j,k)}). \quad (3)$$

The gate input concatenates current and next states from both models with the normalized token position:

$$g_t = [\tilde{h}_t^{(1)}; \tilde{h}_{t+1}^{(1)}; \tilde{h}_t^{(2)}; \tilde{h}_{t+1}^{(2)}; p_t]. \quad (4)$$

Layer normalization prevents a high-magnitude layer from dominating the uniform mean.

3.4 Sequence-Level FiLM Conditioning

Inspired by [Basani and Chen \(2026\)](#), we compute the aggregated features from each model’s token log-likelihood sequences: mean, standard deviation, variance, skewness, kurtosis, first-difference statistics, second-order variance and entropy, and lag autocorrelation, as well as DFT total energy ([Luo et al., 2026](#)). We also append cross-model features including covariance, correlation, cosine similarity, and a Binoculars score ([Hans et al., 2024](#)).

Let the resulting global vector be s . Inspired by FiLM ([Perez et al., 2018](#)), PAWN++ computes

$$\hat{s} = \text{LayerNorm}(s), \quad (5)$$

$$[\gamma, \beta] = W_{\text{film}} \hat{s} + b_{\text{film}}, \quad (6)$$

$$F' = (1 + \gamma) \odot F + \beta. \quad (7)$$

γ and β are initialized with zeros, so the initial model satisfies $F' = F$ and begins as the unconditioned detector. The resulting affine conditioning refines the complete representation without replacing its token pathway.

Figure 2 shows the high-level schema of the complete PAWN++ model. Both LLMs remain frozen; only the lightweight networks are optimized.

4 Experimental Setup

4.1 Data

We use the MAGE cross-domain, cross-model split ([Li et al., 2024](#)), reversing the original labels to match PAWN: 0 denotes human and 1 denotes machine. The full training, validation, and test sets contain 319,071, 56,792, and 60,743 examples, respectively. The test set is nearly balanced (30,478 human and 30,265 machine). Development ablations use balanced subsets of 5,000 training and 2,000 validation examples; every reported sampled-regime score is still measured on the full 60,743-example test set.

OOD evaluation uses MAGE’s unseen-domain-and-model GPT split and its paraphrasing-attack

Method	RAID	HC3		HC3-Chinese		Beemo	
	R@5%FPR	AUROC	Avg. R.	AUROC	Avg. R.	H vs. M	H vs. G
RADAR	70.89	54.59	64.38	35.44	47.93	51.41	60.16
Binoculars	80.61	99.30	91.01	97.35	93.29	83.90	78.15
PAWN++ (ours)	91.83	99.56	93.91	97.98	91.96	81.85	77.78

Table 3: One shot transfer of MAGE trained checkpoints. RAID reports average recall at a five percent false positive budget over all generator and decoding groups, HC3 reports AUROC and average recall, and Beemo reports AUROC on human versus machine and human versus GPT-4o-edited text. Full per setting tables are in Appendix B.

split. These testbeds isolate transfer to a new generator and robustness to post-editing.

4.2 Models and Training

The primary and secondary frozen probes are Llama-3.2-1B-Instruct and Llama-3.2-1B (Grattafiori et al., 2024). Architecture and optimization hyperparameters are listed in Appendix A. The best checkpoint is selected by validation AUROC.

The project includes single-model PAWN and fine-tuned RoBERTa-base baselines. For the final comparison tables, we instead use the original PAWN article’s published MAGE baselines and report AUROC, class-specific recall, and their arithmetic mean.

5 Results on MAGE

5.1 Ablation Study

Table 1 reports the ablations trained on a balanced subset of 5,000 MAGE training examples and selected using a balanced 2,000-example validation subset. All configurations use the same subsets and are evaluated on the full 60,743-example test set. The original PAWN reaches 0.8753 AUROC with the Instruct probe (Llama-3.2-1B-Instruct) and 0.8953 with the Base probe (Llama-3.2-1B). Adding the second model’s token metrics raises AUROC to 0.9126. XPPL and standard hidden-state fusion reduce this score to 0.9096 and 0.9045, respectively, whereas uniform layer fusion raises it to 0.9288. The complete PAWN++ reaches the highest AUROC, 0.9309, after adding sequence-level conditioning. Thus, second-model token metrics provide the initial cross-model gain, uniform fusion is the strongest architectural recovery, and global features add a smaller final improvement.

5.2 In-Distribution MAGE Test

Table 2 compares PAWN++ with the MAGE Testbed 4 baselines reported in the original PAWN

article (Miralles-González et al., 2026). PAWN++ reaches 98.36% AUROC and the highest average recall in this comparison (95.16%). Its AUROC is competitive with the strongest reported PAWN and Longformer results, while its class recalls are more balanced than the Longformer baseline.

We additionally evaluate PAWN++ on the two in-distribution settings reported in Basani and Chen (2026): Testbed 2 uses arbitrary domains with the fixed generator GPT-J, while Testbed 3 uses the fixed WritingPrompts (WP) domain with arbitrary generators. PAWN++ reaches 99.94% AUROC and 98.89% average accuracy on Testbed 2, and 99.81% AUROC and 98.72% average accuracy on Testbed 3. In both settings, it exceeds the reported DivEye AUROC and average accuracy.

5.3 Out-of-Distribution Evaluation

Table 8 places the PAWN++ OOD results beside the original PAWN article’s MAGE Testbeds 7 and 8 baselines (Miralles-González et al., 2026). On the unseen-domain-and-model testbed, PAWN++ has the highest AUROC and average recall.

Paraphrasing produces a qualitatively different outcome. PAWN++ has 79.13% AUROC, which is below RADAR-FT’s 84.00%, confirming that paraphrase robustness remains unresolved. Looking more closely PAWN++ retains high human recall (0.9304) but AI recall falls to 0.4625, indicating that many paraphrased generations cross the fixed decision boundary. Its 79.13% AUROC trails RADAR-FT (84.00%), although it exceeds both original PAWN variants on this metric. The attack therefore removes or weakens the likelihood regularities on which the richer model relies; architecture alone does not substitute for paraphrase-aware training and threshold calibration.

6 One Shot Transfer

A detector that needs retraining for every target distribution cannot serve as shared infrastructure.

We therefore apply the MAGE trained PAWN++ checkpoint to three external benchmarks exactly as it is, with no weight updates, and decision thresholds follow each benchmark’s published evaluation protocol. Table 3 summarizes the transfer, and Appendix B carries the complete comparisons.

6.1 RAID

RAID evaluates detectors by recall at a fixed five percent false positive rate across eleven generators, four decoding configurations, and adversarial attacks (Dugan et al., 2024). Table 11 places PAWN++ in the published comparison. With per domain thresholds held at the target budget, PAWN++ reaches 91.83 percent average recall across all generator and decoding groups, ahead of every detector in the comparison that was not trained on RAID, including Binoculars at 80.61 and the commercial systems. The margin is widest exactly where zero shot methods break, on sampling with repetition penalty, where PAWN++ keeps recall above 92 percent while Binoculars falls to 0.6 percent on non chat models.

6.2 Cross-Dataset Evaluation on HC3

HC3 is a bilingual question-answering benchmark that pairs human and ChatGPT answers to the same questions across multiple domains. Its English and Chinese releases provide distinct cross-lingual evaluation settings for machine-generated-text detection (Guo et al., 2023).

We evaluate PAWN++ trained on the full MAGE training set directly on both HC3 versions, without adaptation. On HC3 English, PAWN++ achieves 99.56 AUROC and 93.91 average recall, the strongest results in Table 10. On HC3-Chinese, it reaches 97.98 AUROC and 91.96 average recall (Table 10), the highest AUROC in the comparison, while Binoculars keeps the best average recall at 93.29.

6.3 Cross-Dataset Evaluation on Beemo

Beemo is a multi-author machine-generated-text benchmark containing human-written responses, outputs from ten instruction-tuned LLMs, expert-edited outputs, and outputs edited by Llama3.1-70B-Instruct or GPT-4o. It covers open-ended generation, rewriting, summarization, and open and closed question answering, making it suitable for evaluating both cross-dataset transfer and robustness to post-generation editing (Artemova et al., 2025).

We evaluate the PAWN++ checkpoint trained on the full MAGE training set directly on Beemo, without fine-tuning or threshold calibration on Beemo. Table 9 adds PAWN++ to the seven binary AUROC comparisons reported by Artemova et al. (2025). PAWN++ reaches 81.85 AUROC on human-written versus unedited machine-generated text and remains strong when Llama- or GPT-edited text is contrasted with human writing (81.82 and 77.78, respectively). Expert editing remains challenging: the human-versus-expert-edited AUROC is 63.95. The near-random Llama-edited-versus-machine score (51.87) further indicates that PAWN++ tends to retain a machine-generated signal after Llama editing, but does not reliably distinguish the edited output from its unedited machine-generated counterpart.

7 WATTLE: A Web Benchmark

Existing benchmarks draw human text from curated sources, while web pages carry boilerplate and noisy register, arrive in many languages, and machine text on the web is produced by the strongest available generators. We build WATTLE (Web Authorship Testbed for Transformations, Languages, and Eras), a benchmark of 41,300 objects in which every machine text is anchored to a real web page and every comparison is source matched. Figure 3 maps the composition, and Appendix C details every cell (Table 6).

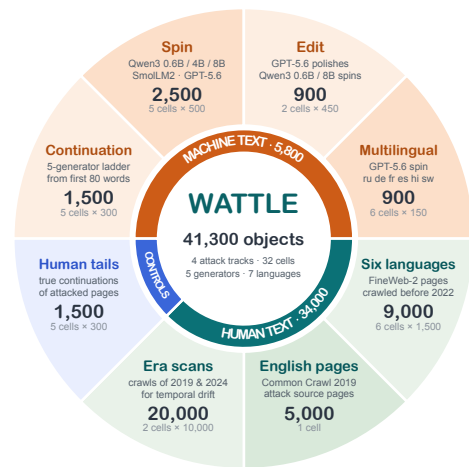


Figure 3: Composition of WATTLE. Four machine tracks anchored to source pages, three human blocks predating the modern generation era, and matched tail controls.

Human gold. All human text predates the post ChatGPT proliferation of machine content. English pages come from a 2019 Common Crawl snapshot (Common Crawl Foundation, 2019), and six further languages (ru, de, fr, es, hi, sw) come from FineWeb-2 shards crawled before 2022 (Penedo et al., 2025). Pages are deduplicated globally, capped per host, and assigned to calibration and test partitions by a salted hash of the URL. Two era scans of ten thousand pages each, drawn from 2019 and 2024 crawls, support a temporal analysis.

Tracks. Machine text is generated from the same source pages along four tracks. *Spin* rewrites the page in the generator’s own words across a capability ladder of Qwen3 0.6B to 8B (Qwen Team, 2025), SmoLLM2 1.7B (Ben Allal et al., 2025), and the frontier model GPT-5.6. *Continuation* extends the page from its first eighty words, with matched human tail controls that pit mid document machine text against the true continuation of the same page. *Edit* has the frontier model lightly polish machine rewrites from a weak and a strong generator (Artemova et al., 2025). *Multilingual* applies frontier spin in all six languages against same language human pages.

Protocol. Generation is one shot, with one fixed prompt per track, fixed sampling, one attempt per seed, and no regeneration of rejected outputs. Every generation passes a single acceptance predicate covering length, refusal filters, language identity, and bounded source overlap. Released cells are subsampled to round sizes by ascending hash of the URL, a rule fixed before any detector was run, so membership cannot leak detector information. The release ships a manifest with per file checksums, full provenance, and complete acceptance statistics (Appendix C).

8 Can Detectors Filter the Web?

8.1 Methods and Protocol

We compare seven frozen detector families: PAWN++, canonical Binoculars with the Falcon-7B pair (Hans et al., 2024; Almazrouei et al., 2023), a probe matched perplexity ratio sharing the Llama probes of PAWN++, Fast-DetectGPT (Bao et al., 2024), a plain log likelihood score, RADAR (Hu et al., 2023), FastText (Joulin et al., 2017), and RoBERTa fine tuned on MAGE (Liu et al., 2019). Every machine text is paired with the human score of its own source page, operating points use each

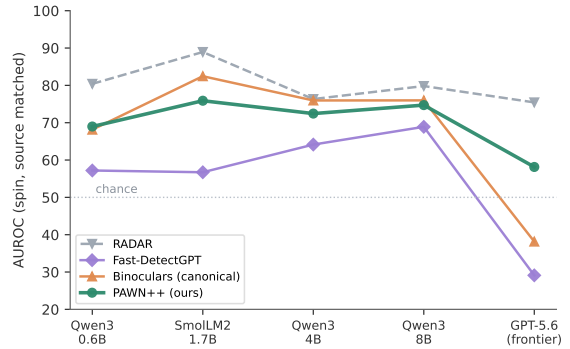


Figure 4: Spin detectability across the generator ladder. Every frozen likelihood detector weakens toward the frontier generator, ratio and curvature scores fall below chance, and PAWN++ stays above chance while degrading gracefully.

method’s one percent threshold from a held out calibration split, uncertainty comes from bootstrap over URL clusters, and method deltas are tested with McNemar on shared pages. A host disjoint recomputation of the English tracks moves AUROC by at most 0.30 points for every method (Appendix C).

8.2 Results

Continuation from open generators is detectable.

PAWN++ reaches 97.9 to 100.0 AUROC per open generator cell with 81 to 99 percent recall at the one percent budget, and matched tail controls confirm the signal survives a mid discourse comparison. Supervised baselines rank the classes well and collapse at the operating point.

The frontier regime breaks frozen detectors.

Rewrites by GPT-5.6 invert likelihood scores. Canonical Binoculars falls to 38.2 AUROC, the probe matched ratio to 45.6, and Fast-DetectGPT to 29.1, all below chance, while frozen PAWN++ stays above chance at 58.1 (Figure 4). The same shift appears in continuation, where every frozen method drops from near perfect scores to between 45 and 70 AUROC on the frontier cell. Detectability is flat across open generator sizes from 0.6B to 14B, and the frontier generator forms a distinct low detectability outlier. No frozen method recalls more than 12 percent of frontier rewrites at the one percent budget.

Calibrated budgets fail to transfer across languages.

Applying the English threshold to six languages breaks every frozen likelihood method somewhere. The probe matched ratio reaches 9.5

Method	Continuation		Spin		Frontier spin
	AUROC	TPR@1%	AUROC	TPR@1%	AUROC
Log likelihood	64.4	6.5%	55.4	0.5%	44.7
Fast-DetectGPT	67.3	26.7%	53.3	2.8%	29.1
FastText	73.5	14.6%	65.8	4.3%	60.9
RoBERTa (MAGE)	87.5	27.5%	73.5	6.9%	67.3
RADAR	94.8	20.3%	80.3	1.5%	75.4
probe matched ratio	94.7	77.9%	77.7	11.8%	45.6
Binoculars	93.0	75.2%	67.9	11.7%	38.2
PAWN++ (ours)	92.0	74.8%	70.1	8.6%	58.1

Table 4: Detection on WATTLE by regime. Source matched paired AUROC and recall at each method’s own calibrated one percent false positive threshold. Continuation and spin pool five generator cells each, and frontier spin is the GPT-5.6 rewrite cell. Multilingual results with backbone coverage conditioning are in Appendix E.

percent false positives on Hindi, a language its backbone officially supports, and 65 percent on Swahili outside its coverage. RADAR and FastText keep their budgets with zero recall. Frozen PAWN++ degrades least, with a worst case of 1.9 percent on supported languages (Appendix E).

Temporal drift. The flagged fraction under the frozen threshold rises from 1.49 percent on the 2019 crawl to 13.44 percent on 2024 (Appendix F), observational evidence of a changing web.

8.3 Training on the Benchmark Inflates Scores and Breaks Transfer

PAWN++ trains a 2.7M parameter head over frozen probes, so supervising it for a new target costs one epoch on a few thousand texts. We train a fresh head on five thousand balanced texts from 2024 crawl pages whose URLs never appear in the benchmark, with the machine side produced by the open generator wave of the same pipeline, and evaluate on the frozen release, which is disjoint from training in URLs and time.

The in protocol numbers look spectacular. Pooled spin rises from 70.1 to 96.4 AUROC, continuation reaches 99.4, the held out frontier cell rises from 58.1 to 96.8, and the six held out languages score 90.5 to 96.1 with false positive rates at or below 2.7 percent. A FastText classifier trained on the same texts reaches only 56.9 AUROC on the frontier cell, so these gains are not a surface artifact that a lexical model could copy.

External evaluation reverses the verdict. On Beemo, machine text produced by a different pipeline with different prompts, the benchmark trained head scores 10 to 26 AUROC points below the frozen head on every comparison, falling from 84.5 to 58.5 on unedited machine text. Against matched human tails the benchmark trained contin-

uation numbers drop from 99.4 to 80.0. The head therefore learned the generation protocol of the benchmark rather than machine text in general, and its benchmark scores overstate real capability by a wide margin. We report the frozen configuration as the honest measure of every detector throughout this paper, and we read this experiment as a warning for the field: a detector trained on a benchmark cannot be certified by that benchmark, and deployment grade evaluation needs generation pipelines that the detector never saw.

9 Conclusion

This work carried a single detector from a curated benchmark into deployment conditions. On MAGE, PAWN++ delivers the most balanced in distribution results in its comparison and the best transfer under domain and generator shift, and the untouched checkpoint holds up across RAID, HC3, and Beemo. The picture changes on WATTLE: statistics that looked reliable for years stop separating classes once the generator is a frontier model, and thresholds calibrated on English quietly lose their meaning in other languages. A closing experiment trains the small head on five thousand texts from the benchmark pipeline and finds that the resulting scores are an illusion: in protocol AUROC jumps to 96.4 while the frozen checkpoint stays ahead by up to 26 points on external machine text. Detection on the web remains open, and progress claims should rest on frozen evaluation against generation pipelines the detector never saw.

Limitations

Our experiments use compact 1B-parameter probe models, so the reported behavior is specific to this compute and model regime. WATTLE covers six non-English languages and uses one fixed prompt

per transformation track and a single frontier-model provider. Its results therefore do not establish generalization to other languages, prompt families, or providers. Generator behavior changes rapidly, so the findings characterize only the documented model versions and evaluation period. The temporal crawl is unlabeled, and changes in the fraction of pages flagged by a detector cannot be interpreted as estimates of AI-generated content prevalence. Multilingual operating points are calibrated on English and depend on the language coverage of the frozen probes. Finally, detector scores provide statistical evidence about text provenance and should not be treated as standalone proof of authorship for an individual document.

Ethics Statement

All human text in WATTLE originates from publicly available web crawls released for research use, Common Crawl ([Common Crawl Foundation, 2019](#)) and FineWeb-2 ([Penedo et al., 2025](#)). We do not intentionally collect private data or target individual authors, and public pages may still contain personal information, which the release mitigates by storing only what reproducibility requires. Pages serve as corpus documents, the release stores only the URLs and text needed for reproducibility, and no attempt is made to identify or profile the people behind them. Machine text is generated by publicly documented models under their respective licenses. Detection scores are intended to support moderation and corpus curation with human oversight, and this work does not develop or release techniques for evading detection. All experiments used shared academic infrastructure with compact 1B parameter probes, keeping the compute footprint modest.

Acknowledgments

This work is an output of a research project (HSE-BR-2025-025) implemented as part of the Basic Research Program at HSE University.

References

P. Akhilesh, Anjali Chennupati, Bhamidipati Prahas, Bommisetty Durga Jasvitha, and Manju Venugopalan. 2025. [AI generated text detection using LORA based fine-tuning of LLM models](#). In *2025 3rd World Conference on Communication & Computing (WCONF)*, pages 1–7. IEEE.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hessel, Julien Launay, Quentin Malartic, et al. 2023. [The falcon series of open language models](#). *Preprint*, arXiv:2311.16867.

Ekaterina Artemova, Jason S. Lucas, Saranya Venktraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. 2025. [Beemo: Benchmark of expert-edited machine-generated outputs](#). In *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)*, pages 6992–7018, Albuquerque, New Mexico. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. [Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature](#). In *The Twelfth International Conference on Learning Representations*.

Advik Raj Basani and Pin-Yu Chen. 2026. [Diversity boosts AI-generated text detection](#). *Transactions on Machine Learning Research*. ArXiv:2509.18880.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. [SmolLM2: When smol goes big – data-centric training of a small language model](#). *Preprint*, arXiv:2502.02737.

Common Crawl Foundation. 2019. The Common Crawl corpus. <https://commoncrawl.org>.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. [The impact of AI-generated text on the internet](#). *Preprint*, arXiv:2604.26965.

George Drayson, Emine Yilmaz, and Vasileios Lampsos. 2025. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29657–29673, Suzhou, China. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. [RAID: A shared benchmark for robust evaluation of machine-generated text detectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Arka Dutta, Utkarshani Jaimini, Utkarsh Bhatt, Sara Shree Muthuselvam, Amitava Das, and Ashiqur Rahman KhudaBukhsh. 2026. A large scale

- social web audit of AI generated text detection systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 677–690.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.
- Sichang Steven He, Calvin Ardi, Ramesh Govindan, and Harsha V. Madhyastha. 2026. [DeGenTWeb: A first look at LLM-dominant websites](#). *Preprint*, arXiv:2605.00087.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 15077–15095.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- 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)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. 2026. [SpecDetect: Simple, fast, and training-free detection of LLM-generated text via spectral analysis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32356–32364.
- Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2026. [Not all tokens are created equal: Perplexity attention weighted networks for AI-generated text detection](#). *Information Fusion*, 125:103465.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.
- Marco Murgia, Diego Reforgiato Recupero, and Georgios Spathoulas. 2026. [A comprehensive evaluation of fine-tuned LLMs as AI-generated text detectors under adversarial and multi-domain conditions](#). *IEEE Access*, 14:66819–66837.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. [FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [FiLM: Visual reasoning with a general conditioning layer](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951.
- Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. 2023. [On the zero-shot generalization of machine-generated text detectors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4799–4808, Singapore. Association for Computational Linguistics.
- Gerrit Quaremba, Elizabeth Black, Denny Vrandečić, and Elena Simperl. 2025. [WETBench: A benchmark for detecting task-specific machine-generated text on Wikipedia](#). *Preprint*, arXiv:2507.03373.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Brian Tufts, Xuandong Zhao, and Lei Li. 2025. [A practical examination of AI-generated text detectors for large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4839–4856, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaowei Zhu, Yubing Ren, Fang Fang, Shi Wang, Yanan Cao, and Li Guo. 2026. [Exons-Detect: Identifying and amplifying exonic tokens via hidden-state discrepancy for robust AI-generated text detection](#). *Preprint*, arXiv:2603.24981.

A Training Hyperparameters

Table 5 lists the architecture and optimization settings shared by every PAWN++ training run in this paper.

Hyperparameter	Value
Maximum sequence length	512
MLP width	256
MLP hidden layers	3
Optimizer	AdamW
Maximum epochs	5
Learning rate	10^{-3}
Weight decay	10^{-2}
Token dropout	0.15
Label smoothing	0.2
Batch size	32
Random seed	42
Checkpoint criterion	Validation AUROC

Table 5: PAWN++ architecture and training hyperparameters.

B Complete Benchmark Results

This appendix reports every published comparison in full for the benchmarks of Sections 5 and 6.

Table 7 reports the two additional in distribution MAGE settings, a fixed generator with arbitrary domains and a fixed domain with arbitrary generators, where PAWN++ exceeds every published baseline including DivEye.

Table 8 places PAWN++ beside the published out of distribution baselines on MAGE testbeds 7 and 8. The unseen domain and model testbed shows the strongest transfer in the comparison, and the paraphrase testbed shows the operating point cost that motivates the benchmark training study of Section 8.3.

Table 9 gives all seven binary Beemo comparisons against the published baselines, covering unedited machine text and machine text refined by experts or stronger models.

Table 10 reports the complete metric sets on HC3 English and Chinese side by side for checkpoints applied without adaptation.

Table 11 expands the RAID comparison into the full recall grid over generator groups, decoding strategies, and repetition penalty, with the published values reproduced from prior work.

C The WATTLE Benchmark

Table 6 breaks the release into its 32 cells. The machine side holds 5,800 texts across four attack tracks, every one generated from a source page that also appears on the human side. The generator ladder spans open weight models from 0.6B to 8B parameters and one frontier model, so the benchmark measures detectors against both accessible and strongest available generation. The human side

holds 34,000 documents, and the five control cells hold the true continuations of the pages attacked in the continuation track.

Sources. English human gold comes from the CC-MAIN-2019-30 snapshot of Common Crawl (Common Crawl Foundation, 2019) with boilerplate removal, global URL deduplication, and per host caps. The six language cells use FineWeb-2 shards crawled before 2022 (Penedo et al., 2025) with 1,500 pages per language. The released human side spans 29,386 unique hosts across its 34,000 pages, so no single site dominates the benchmark. A salted URL hash assigns every document to the calibration or test partition.

Generation prompts. Each track uses one fixed prompt for every generator, with default sampling for the frontier model and temperature 1.0 with top-p 0.95 for the open weight ladder. The spin track uses:

Spin Prompt.

Rewrite the following web page in your own words. Keep the same topic, the same language, roughly the same length, and the same level of detail. Output only the rewritten text, with no preamble or commentary.

The continuation track receives the first eighty words of the page and a length target derived from the original:

Continuation Prompt.

Continue the following web page. Keep the same topic, language, style and register. Write about {target words} more words. Output only the continuation, with no preamble or commentary.

The edit track passes a machine rewrite back to the frontier model:

Edit Prompt.

Lightly edit the following text to improve its flow and wording. Fix awkward phrasing where needed, but keep the meaning, structure, language, and approximately the same length. Output only the edited text, with no preamble or commentary.

The multilingual track reuses the spin prompt, and the phrase about keeping the same language constrains the output to the language of the source page.

Acceptance and release. The acceptance predicate requires at least 400 characters, passes refusal and preamble filters, verifies language identity, and bounds source overlap at 0.5 for rewrites within a length ratio of 0.6 to 1.6, while edits must keep

Track	Cells	Sources and generators	Per cell	Total
<i>Machine text, anchored to source pages</i>				
Spin	5	Qwen3 0.6B/4B/8B, SmolLM2 1.7B, GPT-5.6	500	2,500
Continuation	5	the same generator ladder	300	1,500
Edit	2	GPT-5.6 over Qwen3 0.6B and 8B rewrites	450	900
Multilingual	6	GPT-5.6 in ru, de, fr, es, hi, sw	150	900
<i>Human text, pre LLM</i>				
English pages	1	Common Crawl 2019	5,000	5,000
Per language	6	FineWeb-2 crawled before 2022	1,500	9,000
Era scans	2	crawls of 2019 and 2024	10,000	20,000
<i>Controls</i>				
Human tails	5	true continuations of source pages	300	1,500
WATTLE total	32			41,300

Table 6: Composition of WATTLE. Every machine text is anchored to a real web page, released cells are selected from accepted generation pools by a score blind hash rule, and human text predates the post ChatGPT era.

Testbed 2: Arbitrary domains, GPT-J					Testbed 3: WritingPrompts, arbitrary models				
Method	AUROC	Human R.	AI R.	Avg. R.	Method	AUROC	Human R.	AI R.	Avg. R.
LogRank	67.00	58.81	63.94	61.38	LogRank	76.00	89.61	56.15	72.88
Entropy	83.00	76.43	76.84	76.64	Entropy	78.00	85.96	60.40	73.18
DetectLLM	72.00	66.36	62.07	64.21	DetectLLM	91.00	88.54	80.77	84.66
FastDetectGPT	59.00	62.31	50.49	56.40	FastDetectGPT	76.00	87.25	54.08	70.67
Binoculars	69.00	60.11	65.22	62.67	Binoculars	77.00	80.80	62.07	71.44
BiScope	93.00	89.62	84.86	87.24	BiScope	94.00	91.78	95.27	93.53
DivEye	97.00	90.63	88.56	89.60	DivEye	99.00	92.22	96.88	94.55
PAWN++ (ours)	99.94	99.24	98.54	98.89	PAWN++ (ours)	99.81	98.97	98.47	98.72

Table 7: Additional MAGE in-distribution results. Baselines, testbed definitions, and DivEye values are from Basani and Chen (2026); PAWN++ is our result in the corresponding MAGE settings.

Testbed 7: Unseen domain and model					Testbed 8: Paraphrasing attacks				
Method	AUROC	Human R.	AI R.	Avg. R.	Method	AUROC	Human R.	AI R.	Avg. R.
Longformer	94.00	52.50	99.14	75.82	Longformer	75.00	52.16	81.73	66.94
Longformer [†]	94.00	88.78	84.12	86.54	Longformer [†]	75.00	88.78	37.05	62.92
RADAR ^{*‡}	87.00	82.41	73.62	78.02	RADAR ^{*‡}	71.00	67.67	59.87	63.77
RADAR-FT	94.00	61.42	98.12	79.77	RADAR-FT	84.00	53.33	92.25	72.79
Binoculars ^{*‡}	88.00	5.51	7.12	17.20	Binoculars ^{*‡}	55.00	6.27	16.00	21.76
LogRank	68.00	83.87	43.95	63.91	BiScope	81.00	48.80	89.79	69.30
Entropy	66.00	74.93	50.18	62.55	DivEye	87.00	69.75	83.22	76.49
DetectLLM	73.00	63.66	67.40	65.53	PAWN (GPT2)	75.00	87.00	39.13	63.06
FastDetectGPT	70.00	93.38	41.50	67.44	PAWN (Llama3-1B)	74.00	77.53	55.87	66.70
BiScope	92.00	86.00	82.58	84.24	PAWN++ (ours)	79.13	93.04	46.25	69.65
DivEye	87.00	69.75	83.22	76.49					
PAWN (GPT2)	94.00	84.12	89.50	86.81					
PAWN (Llama3-1B)	97.00	78.87	96.75	87.81					
PAWN++ (ours)	98.42	93.04	95.38	94.21					

Table 8: OOD MAGE results. Original baselines are from Miralles-González et al. (2026); additional methods are from Basani and Chen (2026); PAWN++ is our result. The additional source reports class accuracies (HumanAcc/MachineAcc) and AvgAcc. * unsupervised; ‡ threshold not selected; † threshold selected for Longformer.

overlap between 0.15 and 0.95. Acceptance statistics ship with the release manifest together with per file checksums, prompts, and generator versions. The split is URL level. Re evaluating the English tracks on test pages whose host never occurs in calibration changes AUROC by at most 0.30 points for every method, so host overlap does not affect the results.

D Per Cell Results and Probe Order Control

Table 12 breaks the pooled web results of Table 4 into representative cells and adds the reversed probe order of Binoculars as a control.

Matched human tail controls give 95.4 to 98.9 AUROC for PAWN++ on open generator continuations, so continuation detection is a signal about the text rather than its position in the document. On

Detector	H vs. M	E vs. M	H vs. E	L vs. M	H vs. L	G vs. M	H vs. G
<i>Training-free detectors</i>							
Binoculars	83.90	76.79	61.24	59.90	79.90	57.93	78.15
Log Probability	87.77	68.68	78.02	56.23	84.52	62.45	81.27
Rank	74.22	57.65	69.11	52.55	72.52	55.10	70.81
Log-Rank	86.91	68.13	77.04	56.04	83.62	62.09	80.33
Entropy	10.36	32.02	17.74	44.01	12.82	38.01	15.28
DetectLLM: Likelihood Log-Rank Ratio	79.68	64.12	69.40	54.71	76.25	59.41	72.83
DetectLLM: Normalized Perturbed Log-Rank	74.12	68.30	68.11	54.52	72.49	50.68	73.84
DetectGPT	72.76	69.37	68.21	61.36	70.14	63.05	69.90
<i>Training-based detectors</i>							
FastText	60.18	56.09	54.24	57.34	53.03	54.47	55.39
RADAR	51.41	50.93	50.80	34.51	62.65	37.96	60.16
MAGE	73.72	60.88	64.09	57.95	67.79	60.59	65.37
AIGC MPU	70.52	70.35	50.66	65.71	57.17	59.66	62.69
OpenAI RoBERTa-base	66.95	66.83	49.13	66.95	50.00	66.95	50.00
OpenAI RoBERTa-large	62.93	63.96	48.76	64.71	48.57	59.67	53.74
PAWN++ (ours)	81.85	70.14	63.95	51.87	81.82	57.13	77.78

Table 9: AUROC scores (%) on Beemo. Bold denotes the best score in each column; underlining denotes the best training-based score. Published baselines and task definitions are from Artemova et al. (2025). PAWN++ and the 50-epoch FastText model are evaluated using checkpoints trained on full MAGE, without adaptation to Beemo. For all single-model training-free detectors, the Qwen2-7B configuration is reported. H=human-written; M=machine-generated; E=expert-edited; L=Llama3.1-70B-Instruct-edited; G=GPT-4o-edited. L and G aggregate three editing prompts.

Method	AUROC	Human R.	AI R.	Avg. R.
RoBERTa-base	98.04	60.87	99.14	80.00
Binoculars	99.30	82.68	99.33	91.01
RADAR	54.59	34.11	94.66	64.38
Fast-DetectGPT	99.41	83.71	99.15	91.43
FastText	68.65	55.69	74.60	65.14
PAWN++ (ours)	99.56	88.29	99.54	93.91

Method	AUROC	Human R.	AI R.	Avg. R.
RoBERTa-base	53.55	30.36	71.79	51.08
Binoculars	97.35	92.37	94.21	93.29
RADAR	35.44	92.33	3.53	47.93
Fast-DetectGPT	97.55	89.34	93.83	91.58
FastText	53.79	98.90	5.02	51.96
PAWN++ (ours)	97.98	87.89	96.04	91.96

Table 10: One shot transfer to HC3 English answers (left) and Chinese answers (right) for checkpoints applied without adaptation.

a shared URL core attacked by all five spin generators, PAWN++ moves from 71.9 to 79.3 across the open ladder and dips to 68.9 on the frontier cell, while the probe matched ratio falls from 84.4 to 58.6. An observational extension on a 2024 generation wave adds a Qwen3-14B point and shows flat detectability from 0.6B to 14B. The largest single drop in the comparison belongs to the probe matched ratio, which loses 26 AUROC points between the shared core and the frontier cell.

E Multilingual Results and Backbone Coverage

Interpreting multilingual results requires the language coverage of each backbone. Llama-3.2 officially supports de, fr, es, and hi among our six, Russian is present in its web pretraining without being listed, and Swahili lies outside both. Falcon-7B covers English and French. RADAR, FastText, and RoBERTa are English only, and RoBERTa is not evaluated on non English text by design. Cells outside a backbone’s coverage measure pretraining exposure rather than detection and are excluded from

headline claims. Figure 5 plots the per language AUROC of every frozen method, and Table 13 adds the operating point detail. Fast-DetectGPT and the probe matched ratio break the transferred budget on Hindi and Swahili, while PAWN++ keeps the realized false positive rate at or below 1.4 percent on every supported language.

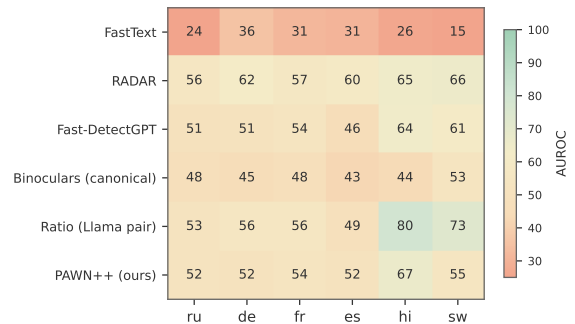


Figure 5: Per language AUROC on frontier multilingual spin. Frontier spin stays close to chance for every frozen method, and cells outside a backbone’s pretraining coverage measure exposure rather than detection.

Table 11: Recall at 5% FPR on RAID (Dugan et al., 2024).

Decoding strategy Repetition penalty	Open Source								Closed Source				Avg.
	Chat (llama-c, mistral-c, mpt-c)				No Chat (mistral, mpt, gpt2)				Chat (c-gpt, gpt4, cohere)		No Chat (cohere, gpt3)		
	Greedy		Sampling		Greedy		Sampling		Greedy	Sampling	Greedy	Sampling	
	No	Yes	No	Yes	No	Yes	No	Yes	No	No	No	No	No
No training on RAID													
R-B GPT2	84.10%	52.30%	77.90%	26.20%	98.60%	44.10%	60.50%	35.40%	70.90%	41.70%	65.10%	52.50%	59.11%
R-L GPT2	79.70%	41.10%	71.40%	19.50%	98.50%	43.00%	67.20%	53.40%	61.40%	34.70%	61.10%	48.60%	56.63%
R-B CGPT	80.20%	63.30%	75.50%	39.30%	53.30%	26.40%	14.90%	1.70%	59.10%	38.10%	46.50%	39.00%	44.78%
RADAR	88.80%	77.40%	85.60%	66.40%	91.80%	63.80%	48.30%	31.80%	81.60%	75.30%	72.20%	67.70%	70.89%
GLTR	89.80%	67.50%	83.90%	38.30%	99.60%	56.90%	44.50%	0.50%	80.70%	54.30%	75.60%	63.70%	62.94%
F-DetectGPT	98.60%	74.50%	96.20%	40.50%	97.80%	56.10%	79.70%	0.60%	96.00%	74.10%	93.80%	86.30%	74.52%
LLMDet	55.50%	30.20%	47.50%	16.50%	74.80%	27.00%	38.40%	3.70%	35.80%	18.50%	40.00%	32.90%	35.07%
Binoculars	99.90%	86.60%	99.70%	60.60%	99.90%	62.30%	72.40%	0.60%	99.20%	92.10%	99.00%	95.00%	80.61%
GPTZero	98.80%	93.70%	98.40%	82.50%	74.70%	34.60%	9.40%	4.80%	92.30%	88.50%	60.60%	53.40%	65.98%
Originality	98.60%	86.30%	97.70%	72.50%	99.90%	64.10%	89.00%	51.20%	96.80%	89.00%	91.70%	85.40%	85.18%
Winston	97.20%	90.10%	96.60%	78.30%	68.20%	49.00%	29.50%	11.30%	96.10%	93.70%	73.20%	68.10%	70.94%
ZeroGPT*	95.40%	80.70%	90.50%	54.90%	85.10%	57.20%	16.00%	0.30%	92.10%	65.80%	83.40%	72.70%	66.18%
PAWN (GPT2)	96.56%	88.57%	93.70%	79.30%	90.53%	84.03%	57.52%	81.56%	90.16%	72.02%	84.34%	75.37%	82.81%
PAWN (Llama3-1b)	98.33%	92.00%	96.26%	81.61%	98.90%	86.85%	52.07%	82.44%	93.10%	73.24%	83.15%	75.04%	84.42%
Longformer	87.89%	86.60%	86.55%	83.43%	86.15%	83.78%	59.62%	83.63%	81.66%	68.24%	64.13%	54.11%	77.15%
RoBERTa	89.89%	98.10%	91.14%	90.26%	85.32%	95.55%	49.14%	89.24%	89.04%	71.97%	68.53%	66.02%	82.02%
RADAR-PTM	88.14%	87.05%	85.80%	82.54%	94.10%	88.66%	49.15%	92.05%	83.51%	67.96%	66.29%	56.80%	78.50%
PAWN++ (ours)	99.39%	98.28%	99.50%	92.82%	91.90%	97.80%	69.03%	95.33%	97.41%	84.57%	92.23%	83.65%	91.83%

Notes. Source values are reproduced from Table 10 of Miralles-González et al. (2026). PAWN++ was evaluated on all 467,985 clean RAID-train rows (Dugan et al., 2024) using a separate threshold for each domain at no more than 5% human FPR. An asterisk marks a source detector that did not achieve the target FPR.

Method	Spin		Continuation		Edit
	Qwen3-8B	GPT-5.6	Qwen3-8B	GPT-5.6	GPT-5.6oq06
FastText	65.7 / 5%	60.9 / 2%	79.6 / 13%	74.1 / 6%	64.2 / 7%
RoBERTa (MAGE)	82.6 / 13%	67.3 / 3%	94.1 / 34%	88.9 / 9%	78.5 / 7%
RADAR	79.8 / 1%	75.4 / 1%	97.4 / 13%	92.7 / 3%	82.2 / 2%
Fast-DetectGPT	68.9 / 3%	29.1 / 0%	97.7 / 77%	45.5 / 1%	62.5 / 2%
probe matched ratio	87.4 / 9%	45.6 / 0%	100.0 / 99%	73.3 / 8%	85.6 / 8%
Binoculars (Falcon)	76.0 / 9%	38.2 / 0%	100.0 / 100%	70.0 / 11%	72.9 / 4%
PAWN++ (ours)	74.7 / 9%	58.1 / 2%	100.0 / 99%	59.3 / 14%	75.2 / 11%

Table 12: Selected cells, AUROC and recall at the one percent budget, source matched. Reversing the Binoculars probe order collapses spin AUROC to between 34 and 45 on every cell, which confirms the canonical orientation.

Method	ru	de	fr	es	hi	sw
FastText	44.1 / 0.0 / 0°	40.2 / 0.1 / 0°	46.3 / 0.0 / 0°	28.7 / 0.0 / 0°	44.8 / 0.1 / 0°	41.2 / 0.0 / 0°
RADAR	56.1 / 0.1 / 0°	61.7 / 0.0 / 0°	55.2 / 0.1 / 0°	59.3 / 0.1 / 0°	63.7 / 0.1 / 0°	65.8 / 0.0 / 0°
Fast-DetectGPT	50.7 / 6.9 / 5	51.2 / 1.5 / 3	54.2 / 2.1 / 2	45.7 / 2.2 / 0	64.3 / 13.5 / 32	60.8 / 7.7 / 10°
probe matched ratio	54.7 / 0.2 / 2	55.2 / 0.3 / 1	57.0 / 0.6 / 0	48.9 / 0.5 / 1	79.1 / 9.5 / 40	72.8 / 65.0 / 88°
Binoculars	48.2 / 1.5 / 1°	45.1 / 0.1 / 0°	48.0 / 0.5 / 0	42.5 / 0.7 / 0°	43.7 / 0.0 / 0°	53.1 / 0.1 / 0°
PAWN++ (ours)	52.1 / 1.3 / 0	51.9 / 0.2 / 0	53.9 / 1.4 / 5	51.7 / 0.5 / 1	67.1 / 0.2 / 0	54.6 / 6.1 / 5°

Table 13: Multilingual frontier spin per language. Each cell reports AUROC, realized false positive percentage at the English one percent threshold, and recall percentage. The circle marks languages outside the backbone’s pretraining coverage. Bold false positive rates mark broken budgets.

F Temporal Analysis

The 2019 scan is flagged at 0.11, 1.49, and 5.81 percent under thresholds calibrated for 0.1, 1, and 5 percent, so the budgets hold on in era human text. The 2024 scan is flagged at 2.64, 13.44, and 18.92 percent. We read the scan as a measure of

threshold stability under crawl era shift rather than as an estimate of machine text prevalence, which would require contemporary authorship labels. The widening between the eras appears at every operating point, and the loosest threshold flags nearly one page in five by 2024.

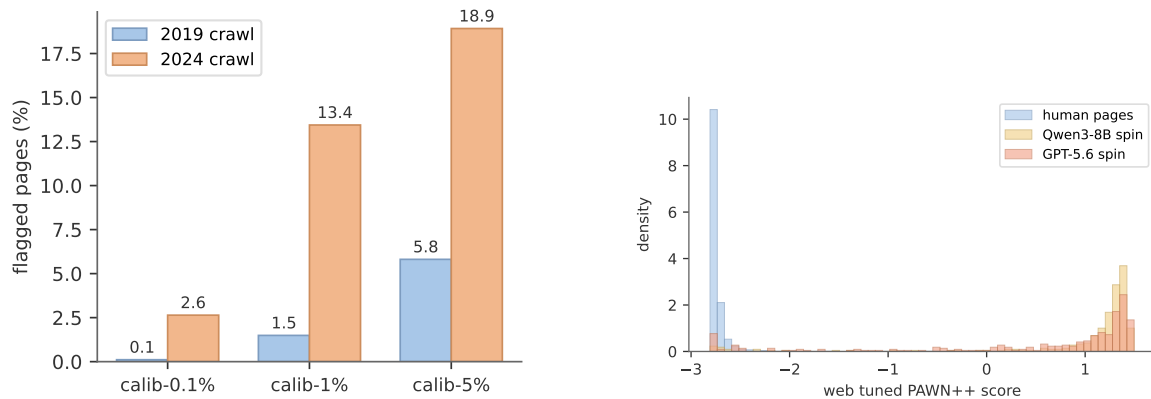


Figure 6: Left: fraction of era scan pages flagged by PAWN++ at three calibrated thresholds, where the 2019 rates match their nominal budgets and the 2024 rates rise far above them. Right: score distributions of the benchmark trained head on human pages, Qwen3-8B spin, and GPT-5.6 spin, whose perfect separation does not survive external machine text (Section 8.3) and is the visual face of benchmark overfitting.

G Selective Prediction

An abstention band built from empirical human score quantiles flags only above the one percent quantile and abstains below the thirty percent quantile, at the cost of abstaining on about thirty percent of human text. For frozen PAWN++ the band lifts continuation recall among decided examples to 100 percent while machine abstention stays at one percent. On spin the same band abstains on about half of the machine texts and decides the rest as human, so abstention rescues continuation and leaves the spin regime open. Audit precision at fifty is measured on a balanced pool and reaches 90 percent on continuation.