Sentence-Level AI Detection for Academic Integrity: A Granular Approach to Reducing False Positives in Mixed-Authorship Submissions

Proofademic Research Team
research@proofademic.ai
Proofademic - proofademic.ai
March 2026

Abstract

We describe a sentence-level classification system for detecting AI-generated text in academic submissions. The system uses a proprietary fine-tuned transformer encoder with a sliding context window of five sentences, trained on ~195k academic documents spanning 14 disciplines and outputs from eight LLMs. Unlike document-level classifiers that return a single score per submission, our system assigns independent confidence values to each sentence, allowing educators to inspect specific flagged passages rather than evaluating a document-wide aggregate. On a held-out evaluation set of 4,200 fully human-authored academic texts, we observe a document-level false positive rate of 0.2%. On a subset of 620 texts by non-native English speakers, the FPR is 0.5% - compared to an average FPR of 61.3% across seven detectors reported in prior work on document-level classifiers applied to similar populations, with at least one detector flagging 97.8% of TOEFL essays as AI-generated (Liang et al., 2023). We also report sentence-level F1 of 92.9% on 2,600 mixed-authorship documents with known ground-truth labels. We release these results alongside a discussion of the training pipeline, including iterative hard negative mining on formulaic and non-native academic prose, and an ablation study over context window sizes.

1. Introduction

Most deployed AI detection systems treat a submission as a single input and return a single probability estimate. This works reasonably well when the document is entirely human-written or entirely machine-generated, which is the scenario that most public benchmarks evaluate. The problem is that this is not how students actually use LLMs.

In practice, the common pattern is partial use: a student writes most of an essay themselves, then uses ChatGPT to generate an introduction, polish a few paragraphs, or draft a literature review section. A

Coursera/Censuswide survey (fieldwork October 2025, published February 2026) found that 80% of students reported AI tools improved their academic performance, while only 20% of educators reported that their university had a formal AI policy in place (Coursera, 2026). The gap is not surprising. What it means for detection is that the real-world input distribution is dominated by hybrid documents, not pure-class ones.

A document-level classifier receiving a 3,000-word essay with two AI-generated paragraphs faces an awkward choice. If its aggregation function is sensitive, it flags the whole document and the educator sees "68% AI probability" with no way to locate the actual AI content. If it's conservative, the signal from two paragraphs gets diluted by 2,500 words of genuine student writing and the system misses it entirely. Either way, the educator gets no useful information about *where* in the text the model is reacting.

The false positive problem compounds this. Liang et al. (2023) showed that seven GPT detectors misclassified over half of TOEFL essays by non-native English speakers as AI-generated - an average FPR of 61.3% on that population, with at least one detector flagging 97.8% of those essays. Turnitin's own documentation reports a 4% sentence-level false positive rate and explicitly states that flagged sentences should be treated as "areas of interest" rather than conclusions (Turnitin, 2023). These are not edge cases. At a mid-sized university processing 50,000 submissions per semester, even a 3% FPR means 1,500 false accusations.

This paper describes a sentence-level detection system built for this setting. The approach is straightforward: instead of one score per document, produce one score per sentence, using a sliding context window to preserve local coherence signals. We report results on a held-out academic corpus with particular attention to FPR disaggregated by native/non-native English writer populations.

2. Related Work

2.1 Document-Level Classifiers

GPTZero, Originality.ai, Copyleaks, and Turnitin's AI writing indicator all operate at the document level, though several now overlay sentence-level highlighting as a post-hoc visualization step. The underlying classifiers are typically transformers trained on binary classification (human vs. machine) using perplexity, burstiness, and distributional features (Mitchell et al., 2023; Gehrmann et al., 2019). GPTZero reports detecting 95.7% of AI text at a 1% document-level false positive rate on the RAID benchmark, with accuracy exceeding 99% when filtered to modern frontier LLMs (GPTZero, 2024). Independent evaluations tell a different story: Dik et al. (2025) found that while GPTZero reliably detected purely AI-generated essays, its reliability on human-authored texts was limited, with false positives across multiple essay lengths. A 2023 evaluation of 14 detectors found that all scored below 80% accuracy and only 5 exceeded 70%, with a systematic bias toward classifying text as human-written (Weber-Wulff et al., 2023).

2.2 Non-Native Writer Bias

Liang et al. (2023) tested seven GPT detectors on 91 TOEFL essays (non-native English speakers) and 88 US eighth-grade essays. The detectors correctly classified the native essays but flagged 61.3% of the TOEFL essays as AI-generated on average, with all seven unanimously misclassifying 19.8% of those essays. The mechanism is well-understood: non-native writers use more predictable vocabulary, shorter sentences, and simpler syntactic constructions. Perplexity-based detectors interpret low perplexity as a machine signal, which creates a direct conflict with the linguistic profile of L2 writers. Vanderbilt disabled Turnitin's AI detector in August 2023 specifically because of this issue (Vanderbilt, 2023).

2.3 Sentence-Level and Sequence Labeling Approaches

Emi and Spero (2024) introduced hard negative mining and synthetic mirror training for reducing false positives in a transformer-based text classifier, achieving low FPRs on high-volume domains like product reviews. Lekkala et al. (2025) framed AI detection as a sentence-level segmentation task, applying CRF layers on top of transformer encoders to detect human-to-AI transitions in hybrid documents. They reported strong gains over zero-shot baselines, particularly on boundary detection. The 2024 ALTA Shared Task (Mollá et al., 2024) established evaluation protocols specifically for sentence-level detection in human-AI collaborative text.

Our work is closest to Lekkala et al. but differs in two ways: we use a classification head rather than CRF decoding (simpler, faster inference at the cost of ignoring global sequence dependencies), and we focus specifically on the academic use case with its attendant fairness requirements around non-native English writers.

3. System Description

3.1 Model Architecture

The classifier is built on a proprietary fine-tuned transformer encoder. We evaluated several pre-trained

encoder architectures during development and selected the one that gave the best F1 on our internal dev set for sentence-level classification. All encoder layers are fine-tuned end-to-end; no layers are frozen.

Input construction works as follows. The document is segmented into sentences using a rule-based splitter with heuristics for in-text citations ("e.g.,", "et al."), abbreviations, and numbered lists common in academic formatting. For each target sentence s_i, we construct an input sequence consisting of s_{i-2}, s_{i-1}, [SEP], s_i, [SEP], s_{i+1}, s_{i+2}. The [SEP]-delimited target sentence is the classification focus; the flanking sentences provide discourse context. At document boundaries, we pad with empty strings.

The classification head is a single linear layer over the pooled encoder representation, producing a scalar logit passed through sigmoid to yield p(AI | s_i, context). We apply temperature scaling (Guo et al., 2017) on a held-out calibration set of 5,000 sentences to produce calibrated confidence scores. Post-calibration, a sentence scored at 0.90 is empirically AI-generated 89.4% of the time (within 1% of stated confidence across all bins above 0.5).

3.2 Training Data

We constructed three training pools:

Human-authored (82k documents). Sourced from arXiv, CORE, Semantic Scholar's ORC, and open-access thesis repositories across 14 disciplines. All documents have verified single-author attribution and were published before November 2022 to avoid ChatGPT-era contamination. We applied automatic language detection and removed non-English documents, leaving 82,341 documents comprising ~48M sentences.

AI-generated (78k documents). For each discipline and format category in the human pool, we generated parallel documents using eight LLMs: GPT-5, Claude Sonnet 4.5, Gemini 2.5 Pro, Llama 4 Maverick, DeepSeek-R1, Mistral Large 2, Qwen 3.5, and Grok 3. Outputs from Qwen 3.5 and Grok 3 were reserved exclusively for evaluation (blind-tested models excluded from the training split) to measure generalization to unseen architectures. We varied prompt strategies across direct generation ("Write a 2000-word essay on..."), outline-then-expand, and rewrite-from-notes to diversify the output distribution. Temperature was sampled uniformly from [0.3, 1.0] for each generation. The final count after deduplication was 78,204 documents.

Mixed-authorship (35k documents). We spliced AI-generated sentences into human-authored documents at four density levels (10%, 25%, 50%, 75%) and four position distributions (front-loaded, back-loaded, middle-concentrated, uniform random). This yielded 35,120 hybrid documents with per-sentence ground truth labels. The splicing deliberately targets realistic patterns: AI introductions grafted onto human analysis, AI literature reviews within human-written methods-and-results structures, and AI transitions bridging human-authored sections.

3.3 Hard Negative Mining

After initial fine-tuning, we run the model over a pool of 200k human-authored sentences held out from training. Every sentence classified as AI with $p > 0.5$ is extracted as a hard negative.

We then retrain from the last checkpoint with the hard negatives mixed into the training set at a 1:3 ratio (one hard negative for every three normal training examples). This cycle repeats five times. The hard negative pool is deliberately enriched with categories that produce elevated FPRs in pilot testing: non-native English academic text (TOEFL/IELTS essay corpora, ESL thesis collections), formulaic disciplinary prose (clinical case reports, legal briefs, patent applications), and highly structured formats (methods sections, lab protocols, statistical results paragraphs).

The effect is substantial. After five mining iterations, the FPR on formulaic academic prose dropped from 8.3% to 1.1%, and the FPR on non-native English text dropped from 9.7% to 1.4%. Without hard negative mining, the model treats low-perplexity, predictable prose as an AI signal regardless of authorship. The mining forces it to distinguish between "predictable because formulaic" and "predictable because machine-generated."

3.4 Synthetic Mirrors

For each human document, we generate a synthetic mirror: an AI-generated text matched in topic, structure, length, and approximate register. The purpose is to prevent the model from learning spurious correlations between subject matter and class labels. Without mirrors, a model trained on human-authored organic chemistry papers and AI-generated psychology papers may learn to classify by discipline rather than by authorship signal. Emi and Spero (2024) demonstrated the effectiveness of this approach for general-domain text; we extend it to a multi-discipline academic setting with format-aware prompting (matching IMRAD structure, section headers, citation density).

4. Experimental Setup

4.1 Evaluation Data

The evaluation set is disjoint from training (no overlapping source documents or authors where attribution was available). It consists of three subsets: (a) 4,200 fully human-authored academic texts, stratified by discipline, document length, and writer demographics, with a 10% random sample manually verified by two annotators (Cohen's kappa = 0.94); (b) 3,800 fully AI-generated texts from all eight LLMs, including the two blind-tested models (Qwen 3.5 and Grok 3); (c) 2,600 mixed-authorship documents with sentence-level ground truth.

The non-native English subset (n=620) was drawn from the human-only pool and includes texts from writers whose first language spans 23 language families. We determined L1 background from author metadata and institutional affiliation where available, supplemented by manual annotation of a 15% sample.

4.2 Metrics

We report four metrics. **Document-level FPR**: percentage of fully human-authored documents where any sentence exceeds the 0.85 confidence threshold (our default production threshold, chosen to optimize for precision over recall in the academic integrity context). **Sentence-level precision, recall, and F1** on the mixed-authorship subset. **FPR by writer population** (native vs. non-native English). **Per-model detection accuracy** on the AI-only subset, including the two withheld model families.

4.3 Baselines

We compare against two baselines: (1) a document-level transformer classifier trained on the same data using the same encoder backbone (single pooled prediction per document, no sentence segmentation), and (2) DetectGPT (Mitchell et al., 2023), a zero-shot method based on log-probability curvature applied at the sentence level. Both baselines use the same 0.85 threshold for FPR calculation.

5. Results

Table 1. Document-level false positive rates and sentence-level detection performance. Best results in each row are bolded.

Metric	Ours	Doc-level	DetectGPT
Doc FPR (all)	**0.2%**	2.9%	14.2%
Doc FPR (native)	**0.6%**	1.8%	8.1%
Doc FPR (non-native, n=620)	**0.5%**	6.4%	31.6%
Sent. precision (mixed)	**94.1%**	n/a	72.3%
Sent. recall (mixed)	**91.7%**	n/a	64.8%
Sent. F1 (mixed)	**92.9%**	n/a	68.3%

The document-level baseline achieves a 2.9% FPR overall, which is competitive with published numbers from commercial tools. But the native/non-native gap is the important number: 1.8% vs. 6.4% for the document-level model, compared to 0.2% vs. 0.5% for ours. Hard negative mining closes this gap from 4.6 percentage points to 0.3 percentage points. DetectGPT's 31.6% FPR on non-native text confirms the known vulnerability of perplexity-based methods to L2 writing patterns.

Table 2. Detection accuracy by source LLM (AI-only subset).

Model	Acc.	Seen
GPT-5	99.9%	Yes
Claude Sonnet 4.5	99.8%	Yes
Gemini 2.5 Pro	99.7%	Yes
Llama 4 Maverick	99.4%	Yes
DeepSeek-R1	99.5%	Yes
Mistral Large 2	99.6%	Yes
Qwen 3.5 *	98.1%	No
Grok 3 *	97.6%	No

** Blind-tested (excluded from training data).*

The ~2 point accuracy drop on the blind-tested models (Qwen 3.5, Grok 3) is expected and consistent with Emi and Spero (2024). It confirms the classifier is learning general distributional properties of machine text rather than memorizing model-specific artifacts, but it also means ongoing retraining is necessary as new LLMs enter the market.

Table 3. Sentence-level F1 by AI content density.

AI Density	Prec.	Rec.	F1
10%	91.2%	86.4%	88.7%
25%	93.6%	90.8%	92.2%
50%	95.1%	93.5%	94.3%
75%	96.3%	95.8%	96.0%

The 10% density case is the hardest and the most practically important. An essay with 10% AI content is three or four sentences in a 40-sentence paper. The system still achieves 88.7% F1 at this density, though recall drops to 86.4%, meaning roughly one in seven AI-generated sentences is missed. In the academic integrity context, this is an acceptable tradeoff: the flagged sentences are enough to initiate a conversation, and the high precision (91.2%) means the flagged sentences are almost always actually AI-generated.

Table 4. Ablation: context window size. FPR measured on full human-only eval set.

Window	Sent. F1	Doc FPR
1 (target only)	87.4%	3.2%
3 (1+1)	90.8%	0.4%
5 (2+2)	**92.9%**	**0.2%**
7 (3+3)	92.7%	0.2%
9 (4+4)	92.4%	0.2%

Performance peaks at a window of 5 (two sentences of context on each side) and plateaus or slightly degrades beyond that. We attribute this to two factors: wider windows start including context from different authorship zones in hybrid documents (adding noise rather than signal), and the increased input length pushes more examples past the encoder's token limit, requiring truncation. The 5-sentence window also has a practical advantage: inference time per sentence remains low enough to process a 10,000-word thesis in under 10 seconds on production hardware.

6. Discussion

The main result is that sentence-level classification with hard negative mining dramatically reduces the FPR gap between native and non-native English writers (from 4.6pp in the document-level baseline to 0.3pp in our system). This matters for institutional deployment. A university that processes submissions from a diverse student body cannot accept a system that disproportionately flags international students; beyond the fairness concern, it creates real legal exposure.

The sentence-level output also changes how the tool gets used in practice. When an educator sees a document score of 68%, they face a binary decision with no supporting evidence. When they see four highlighted sentences in an otherwise clean document, they can point to specific passages and ask the student about them. This shifts the workflow from "the tool says you cheated" to "can you walk me through how you wrote these sections?" - which is closer to how academic integrity conversations should work.

The ablation over window sizes (Table 4) has a practical implication worth noting. The target-only model (window=1) achieves 87.4% F1, which is already usable. Most of the gain comes from adding a single sentence of context on each side (window=3, F1=90.8%). The jump from 3 to 5 adds another 2.1 points, and after that returns diminish. For deployments with tight latency budgets, a window of 3 is a reasonable tradeoff.

6.1 Limitations

Several caveats apply. The mixed-authorship evaluation uses synthetic splicing, not real student hybrid submissions. Real students don't insert AI paragraphs at random positions; they edit, paraphrase, and blend AI output with their own writing in ways that our splicing methodology doesn't fully capture. We expect real-world sentence-level F1 to be lower than the 92.9% reported here, particularly for students who heavily edit AI-generated drafts.

We evaluated on English only. The model has no multilingual capability and we make no claims about performance on non-English academic text.

AI humanizer tools (adversarial paraphrasing) are a known and growing challenge. We did not evaluate against humanizer-modified text in this paper. Prior work shows paraphrasing attacks can dramatically reduce detection accuracy: Krishna et al. (2023) reported DetectGPT accuracy dropping from 70.3% to 4.6% at a 1% FPR under DIPPER paraphrasing, with similar magnitudes reported across other detectors and humanizer tools (Emi and Spero, 2024). This is an active area of work for us.

Finally, the 0.85 threshold used for FPR calculation is a production default, not a universal optimum. Institutions with different risk tolerances will need to adjust this threshold. A lower threshold catches more AI content but increases false positives; a higher threshold is more

conservative. We provide per-threshold FPR curves to institutional partners to support this calibration.

7. Conclusion

We described a sentence-level AI detection system for academic integrity applications. The system produces per-sentence confidence scores using a proprietary transformer encoder with a 5-sentence sliding context window, trained on ~195k academic documents with iterative hard negative mining.

The key numbers: 0.2% document-level FPR overall, 0.5% for non-native English writers (compared to 6.4% for an equivalent document-level classifier and 31.6% for DetectGPT on the same population), 92.9% sentence-level F1 on mixed-authorship documents, and 97.6-99.9% accuracy across eight LLM families including two blind-tested models excluded from training data.

The broader point is that sentence-level granularity is not just a technical improvement over document-level classification - it changes the operational model for how detection tools are used in academic settings. When the output is specific and interpretable, the tool supports evidence-based conversations instead of binary verdicts. That seems like the right direction for this technology.

References

1. Coursera & Censuswide. (2026). *AI in Higher Education Report*. Fieldwork conducted October 15-23, 2025; report published February 25, 2026.
2. Dik, S., Erdem, O., & Dik, M. (2025). Assessing GPTZero's accuracy in identifying AI vs. human-written essays. Stanford SCALE Initiative. arXiv:2506.23517.
3. Emi, B., & Spero, M. (2024). Technical report on the Pangram AI-generated text classifier. arXiv:2402.14873.
4. Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. *Proceedings of ACL 2019 System Demonstrations*.
5. GPTZero. (2024). AI detection benchmarking: Accuracy, transparency and fairness. Technical documentation.
6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of ICML 2017*.
7. Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Proceedings of NeurIPS 2023*.
8. Lekkala, S. T., Yadagiri, A., Pakray, P., Chunka, C., & Vardhan, M. S. (2025). Fine-grained detection of AI-generated text using sentence-level segmentation. arXiv: 2509.17830.
9. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779.
10. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *Proceedings of ICML 2023*.
11. Mollá, D., Xu, Q., Zeng, Z., & Li, Z. (2024). Overview of the 2024 ALTA Shared Task: Detect automatic AI-generated sentences for human-AI hybrid articles. arXiv:2412.17848.
12. Turnitin. (2023, June 14). Understanding the false positive rate for sentences of our AI writing detection capability. Turnitin Blog.
13. Vanderbilt University. (2023, August 16). Guidance on AI detection and why we're disabling Turnitin's AI detector. Brightspace Support.
14. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., et al. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(26).

Disclosure: *This work was conducted by the Proofademic research team. Proofademic is a commercial AI detection platform. The evaluation was performed on the specific datasets and conditions described above; production performance will vary with document characteristics, domain, and other factors not controlled for in this study.*

Correspondence: *research@proofademic.ai*