

HybridDetect: Detecting Partial AI Contamination in Student Essays via Multi-Class Classification

Daniel Dan
School of Applied Data Science
Modul University
Vienna, Austria
ORCID: 0000-0002-7251-7899
daniel.dan@modul.ac.at

Minoru Nakayama
Interdisciplinary Mathematical Sciences
Meiji University
Tokyo, Japan
ORCID: 0000-0001-5563-6901
nakayama@meiji.ac.jp

Abstract—The proliferation of large language models (LLMs) in higher education has created an urgent need for tools that detect AI-generated content. While existing detectors such as GPTZero and Turnitin operate as binary classifiers distinguishing fully human from fully AI-generated text, they fail to address the increasingly common scenario where students partially incorporate AI-generated content into their writing. We introduce HybridDetect, a multi-class classification framework that detects varying levels of AI contamination in student essays across seven granularity levels: 0%, 10%, 25%, 50%, 75%, 90%, and 100% AI content. Our approach combines a fine-tuned RoBERTa model for semantic analysis with an XGBoost classifier trained on 36 handcrafted linguistic features, integrated through weighted soft voting. We construct a synthetic dataset of 50,001 hybrid essays by blending sentences from approximately 487,000 human-written and 487,000 AI-generated essays at controlled contamination levels. On a held-out test set of 7,501 essays, the ensemble achieves 79.9% accuracy and 0.798 macro F1-score on the 7-class task, with 0.876 F1-score on binary classification ($\geq 50\%$ AI). Results reveal that semantic features captured by RoBERTa are essential for fine-grained detection (0.798 macro F1), while surface-level linguistic features alone achieve only 0.394 macro F1, indicating that AI contamination manifests primarily in semantic patterns rather than stylistic markers. These findings have direct implications for academic integrity policies, suggesting that institutions should adopt graduated response frameworks rather than binary pass/fail judgments.

Index Terms—AI detection, hybrid writing, academic integrity, large language models, RoBERTa, natural language processing, higher education

I. INTRODUCTION

THE rapid adoption of large language models (LLMs) such as ChatGPT, Claude, and Gemini in educational settings has fundamentally altered the landscape of academic integrity. Students increasingly use these tools not to generate entire assignments from scratch, but to selectively enhance portions of their work: paraphrasing paragraphs, generating supporting arguments, or polishing conclusions. This partial integration of AI-generated content, which we term *hybrid writing*, represents a significant challenge that current detection tools are ill-equipped to address.

Existing AI content detectors operate as binary classifiers, categorizing text as either “human-written” or “AI-generated” [1], [2]. This binary framing fails in two important

respects. First, it does not capture the reality of how students actually use AI tools, as writing assistants rather than whole-sale replacements. Second, it forces institutions into all-or-nothing judgments that cannot distinguish between a student who used AI to polish one sentence and one who submitted entirely AI-generated work.

The educational implications are substantial: academic integrity policies increasingly recognize degrees of AI assistance [3], yet binary tools flag a student who used AI for 10% of an essay the same as one who submitted 100% AI-generated work, creating both unfairness and enforcement gaps.

We address this gap with **HybridDetect**, a multi-class classification framework that estimates the degree of AI contamination in student essays. Rather than a binary label, HybridDetect assigns texts to one of seven contamination levels: 0%, 10%, 25%, 50%, 75%, 90%, and 100% AI content. This granularity enables institutions to calibrate responses proportionally, from no action at low contamination to formal review at high contamination.

Our contributions are threefold:

- We formulate AI content detection as a multi-class contamination estimation problem, moving beyond the binary classification paradigm that dominates current approaches.
- We construct a synthetic dataset of 50,001 hybrid essays at seven controlled contamination levels, providing a reproducible benchmark for this task.
- We analyze the relative contribution of semantic and surface-level signals by combining fine-tuned RoBERTa with an XGBoost model on linguistic features, showing that semantic representations are critical for fine-grained contamination detection whereas surface-level features alone are insufficient and add no statistically significant signal when ensembled.

II. RELATED WORK

A. AI-Generated Text Detection

Academic detection approaches include DetectGPT [2], which uses perturbation-based log-probability curvature analysis, and watermarking schemes [4] that embed statistical signals during generation.

Several studies have evaluated the reliability of binary detectors. Liang et al. [5] found that detectors exhibit bias against non-native English writers, while Sadasivan et al. [6] demonstrated that paraphrasing attacks can significantly degrade detection accuracy. Weber-Wulff et al. [7] conducted a systematic evaluation of 14 detectors, finding highly variable performance across domains and text lengths. Feature-based detectors using handcrafted stylometric and error-related cues have also been explored; Gryka et al. [8], for instance, detect LLM-generated emails with a random-forest classifier and find that spelling- and editing-correctness features carry measurable discriminative signal.

B. Commercial Detection Tools

A growing ecosystem of commercial tools serves educational use. GPTZero [1] applies perplexity and burstiness analysis with optional sentence-level highlighting; Turnitin [9] reports an overall AI-likelihood percentage within its similarity report; and Winston AI [10] and Copyleaks [11] offer comparable binary services. Adjacent tools such as QuillBot, marketed primarily as AI paraphrasers, bundle a binary “AI detector” without disclosing model details or quantifying the proportion of AI content.

Crucially, all of these tools frame detection as binary, sometimes with a confidence score, and none produce a quantified estimate of overall contamination. A 70% score from GPTZero means “we are 70% certain this is AI-generated,” not “70% of this text is AI-generated.” This distinction is central: the former measures classifier uncertainty, the latter, which our work addresses, measures content composition. To the best of our knowledge, no existing tool, commercial or academic, performs multi-class contamination-level estimation as formulated here.

C. Hybrid and Partially AI-Generated Text

The challenge of detecting partially AI-generated content has received limited attention. Dugan et al. [12] studied human ability to locate the boundary where machine-generated text begins, finding that people struggle to detect the transition, particularly for more advanced models.

In an early, unpublished exploration, Cutler et al. [13] studied automatic detection of boundaries in mixed human-machine text, though their work predates modern LLMs and focused on older generation models. More recently, researchers have begun to examine mixed-authorship scenarios [14], but the problem of quantifying the degree of AI contamination in a text remains largely unexplored.

D. AI in Education and Academic Integrity

The educational community has responded to LLMs with both concern and pragmatism. Cotton et al. [3] and Baidoo-Anu and Ansah [15] argue that institutions need nuanced policies, beyond outright bans, supported by detection tools that enable graduated responses. Examining usage, Smolansky et al. [16] report that educators and students alike perceive generative AI as significantly impacting assessment, while

Perkins et al. [17] propose the AI Assessment Scale (AIAS), a five-level framework whose graduated levels of permitted AI assistance detection tools should align with.

Beyond detection, generative AI is reshaping the full instructional lifecycle, from course design to assessment and feedback, raising parallel integrity and quality concerns [18]. Our work bridges these research areas by developing a detection system explicitly designed to estimate contamination levels, directly supporting the graduated response frameworks advocated by educational researchers.

III. METHODOLOGY

A. Problem Formulation

We formulate hybrid AI detection as a multi-class classification problem. Given a text x , we aim to predict its contamination class $c \in C = \{0, 10, 25, 50, 75, 90, 100\}$, where c represents the approximate percentage of sentences in x that originate from an AI source. Formally, we seek a function $f : \mathcal{X} \rightarrow C$ that minimizes the expected classification error over the joint distribution of texts and contamination levels.

We additionally define a binary task: classifying texts as predominantly AI ($c \geq 50$) or predominantly human ($c < 50$), which directly supports institutional decision-making.

B. Dataset Construction

1) *Source Data*: We use the AI vs Human Text dataset [19], which contains approximately 487,000 human-written essays and 487,000 AI-generated essays covering diverse topics and writing styles. Human essays are sourced from authentic student submissions and web corpora, while AI essays are generated by state-of-the-art language models.

2) *Hybrid Essay Generation*: To construct hybrid essays with controlled contamination levels, we employ a sentence-level blending procedure. For each target contamination level $c\%$, we:

- 1) Randomly select a human essay h and an AI essay a from the source data.
- 2) Tokenize both essays into sentences, yielding sentence sequences $S_h = (s_1^h, s_2^h, \dots, s_m^h)$ and $S_a = (s_1^a, s_2^a, \dots, s_n^a)$.
- 3) Determine the target number of sentences k from the shorter essay.
- 4) Randomly select $\lfloor k \cdot c/100 \rfloor$ positions to replace with AI sentences.
- 5) Replace human sentences at selected positions with corresponding AI sentences.

This procedure preserves approximate essay length while controlling the proportion of AI-generated content. For the boundary cases: 0% contamination uses only human sentences, and 100% uses only AI sentences.

We adopt the fraction of replaced sentences as an operational definition of contamination because it is a precise, reproducible control variable yielding exactly balanced classes. This approximates rather than equals the true AI proportion: since sentences differ in length and semantic weight, the

replaced-sentence count need not match the proportion of AI-generated tokens. Random human–AI pairing is likewise a simplification that does not reflect how students paraphrase or edit AI output. We revisit both assumptions in Section VI.

We generate 7,143 essays per contamination level, yielding a balanced dataset of 50,001 essays. This per-class count substantially exceeds the sample sizes typically required for stable supervised fine-tuning and guarantees equal representation of all seven levels during training and evaluation.

C. Feature Engineering

We extract 36 handcrafted linguistic features in five groups. **Basic statistics** (10): word, character, and sentence counts and their first and second moments (mean, standard deviation, min, max) at word and sentence level. **Vocabulary richness** (4): type-token ratio, hapax legomena ratio, unique word ratio, and the index $V/\sqrt{N}$, where V is the number of distinct word types and N the total tokens. **Punctuation** (12): normalized frequencies of commas, periods, question and exclamation marks, semicolons, colons, quotation marks, parentheses, dashes, digits, uppercase characters, and whitespace. **Readability** (4): average syllables per word, long-word (>8 chars) and short-word (≤ 3 chars) ratios, and a Flesch Reading Ease approximation. **Structural markers** (6): paragraph and line counts, transition-word density, count of AI-typical phrases (e.g., “it is important to note”), and binary indicators for capitalization and terminal punctuation. These features capture known stylistic differences between human and AI writing, such as AI’s tendency toward more formal vocabulary, higher transition-word density, and more uniform sentence structure [20].

D. Classification Models

1) *RoBERTa Fine-Tuning*: We fine-tune RoBERTa-base [21] for 7-class sequence classification. RoBERTa’s bidirectional transformer architecture captures contextual semantic patterns that distinguish varying contamination levels. Input essays are truncated to 512 tokens; approximately 36% of essays in our corpus exceed this limit, so for those documents the model observes only the leading portion, a constraint we revisit in Section VI. The classification head consists of a linear projection from the [CLS] token representation to 7 output logits.

Training uses the AdamW optimizer with learning rate 2×10^{-5} , weight decay 0.01, warmup ratio 0.1, and effective batch size 32 (batch size 8 with gradient accumulation factor 4). We train for 3 epochs with early stopping on validation macro F1 (patience 2).

2) *XGBoost on Linguistic Features*: We train an XGBoost [22] gradient boosted tree classifier on the 36 scaled linguistic features. Hyperparameters include 500 estimators (with early stopping at patience 20), maximum depth 8, learning rate 0.05, subsample ratio 0.8, and column sampling ratio 0.8.

3) *Ensemble via Weighted Soft Voting*: We combine both models through weighted soft voting. For a text x , let $p_R(c|x)$ and $p_X(c|x)$ denote the class probability estimates from RoBERTa and XGBoost, respectively. The ensemble prediction is:

$$\hat{c} = \arg \max_{c \in C} [w_R \cdot p_R(c|x) + w_X \cdot p_X(c|x)] \quad (1)$$

where w_R and w_X are learned weights optimized via grid search over the validation set by maximizing macro F1 score.

IV. EXPERIMENTS

A. Data Split and Evaluation Protocol

The 50,001 essays are split into training (70%, $n = 35,000$), validation (15%, $n = 7,500$), and test (15%, $n = 7,501$) sets using stratified sampling to ensure balanced class representation. All model selection and hyperparameter tuning are performed on the validation set; test set results are reported only once.

B. Evaluation Metrics

We report accuracy, macro-averaged F1-score, weighted F1-score, and per-class F1-scores for the 7-class task. For the binary task ($\geq 50\%$ AI vs. $< 50\%$ AI), we report F1-score and accuracy. Macro F1 is our primary metric as it treats all contamination levels equally regardless of class size.

C. Implementation Details

All experiments run on an Apple M2 Pro/Max (64 GB RAM, 12 cores) with Metal Performance Shaders acceleration; RoBERTa fine-tuning uses the HuggingFace Transformers library [23] and XGBoost the histogram-based tree method. The full pipeline completes in about 3.5 hours, dominated by RoBERTa fine-tuning.

V. RESULTS

A. Overall Performance

Table I presents the performance of all three models on the held-out test set.

TABLE I
PERFORMANCE COMPARISON ON THE 7-CLASS HYBRID DETECTION TASK.

Model	Acc.	Macro F1	Wt. F1	Bin. F1
RoBERTa-base	0.798	0.798	0.798	0.875
XGBoost	0.394	0.394	0.394	0.718
Ensemble	0.799	0.799	0.799	0.876

RoBERTa achieves 79.8% accuracy and 0.798 macro F1, demonstrating strong 7-class discrimination. The ensemble provides a marginal improvement of +0.0004 macro F1 over RoBERTa alone, with the optimal weight configuration of $w_R = 0.70$, $w_X = 0.30$. For binary classification ($\geq 50\%$ AI), the ensemble reaches 0.876 F1-score.

To test whether the ensemble gain is meaningful, we computed 95% confidence intervals via paired bootstrap resampling (5,000 replicates over the test set). The macro F1

interval is $[0.790, 0.807]$ for the ensemble and $[0.789, 0.807]$ for RoBERTa, and the interval for their paired difference, $[-0.003, +0.003]$, comfortably includes zero. A McNemar test on per-sample correctness confirms this ($p = 0.87$). We therefore report the ensemble for completeness but treat it and RoBERTa as statistically equivalent: on this dataset the linguistic-feature model contributes no significant additional signal once semantic features are present.

Because contamination is an inherently ordinal quantity, we additionally report mean absolute error (MAE) on the percentage scale, which penalizes large level errors more than small ones. Both RoBERTa and the ensemble attain an MAE of 13.3 percentage points, and 86.3% of predictions fall within one contamination level of the ground truth, indicating that most errors are to neighboring classes rather than gross misestimates.

The XGBoost model achieves only 39.4% accuracy on the 7-class task, well above the 14.3% random baseline but substantially below RoBERTa. However, its binary F1 of 0.718 indicates that linguistic features can partially distinguish predominantly human from predominantly AI text, even if they cannot resolve fine-grained contamination levels.

B. Per-Class Analysis

Table II shows F1-scores broken down by contamination level.

TABLE II
PER-CLASS F1 SCORES BY AI CONTAMINATION LEVEL (ENSEMBLE MODEL).

Level	RoBERTa	XGBoost	Ensemble
0% AI	0.647	0.404	0.641
10% AI	0.898	0.403	0.898
25% AI	0.851	0.362	0.853
50% AI	0.899	0.432	0.899
75% AI	0.812	0.339	0.813
90% AI	0.858	0.425	0.861
100% AI	0.621	0.389	0.624

The results reveal a distinctive pattern: intermediate levels (10–90%) are classified with substantially higher accuracy (F1 0.81–0.90) than the boundary classes of 0% and 100% (F1 0.62–0.65). The primary error (Fig. 1) is confusion between 0% (pure human) and 100% (pure AI) essays: 344 of 1,071 pure-human essays are misclassified as pure AI, and 365 of 1,072 pure-AI as pure human.

This confusion between pure classes is an artifact of our sentence-level blending: since both 0% and 100% essays consist entirely of sentences from a single source, they lack the distributional mixing signal that makes intermediate classes distinctive.

C. Feature Importance Analysis

Fig. 2 presents the XGBoost feature importance rankings. The most informative linguistic features are structural markers (terminal punctuation, capitalization patterns), vocabulary richness measures (hapax ratio, type-token ratio), and

document structure (line count, paragraph count). Notably, AI-specific markers such as transition word ratio and AI-phrase count rank lower than expected, suggesting that surface-level AI indicators are insufficient for contamination-level estimation. This is consistent with their nature: such lexical heuristics are brittle and model-specific, and are likely to lose discriminative value as language models evolve, which makes their low importance unsurprising rather than incidental.

D. Semantic vs. Linguistic Features

The dramatic gap between RoBERTa (0.798 macro F1) and XGBoost (0.394) indicates that level-wise contamination manifests primarily in semantic rather than surface-level patterns (interpreted further in Section VI). The two models agree on only 40.4% of test predictions. When they agree, accuracy rises to 85.1%, suggesting high confidence in shared predictions. On the 59.6% of disagreements, RoBERTa is correct 76.3% of the time versus XGBoost’s 8.4%, confirming the dominance of semantic features for this task.

E. Ensemble Weight Sensitivity

Across nine weight configurations ($w_R \in [0.30, 0.80]$), ensemble performance is robust for $w_R \geq 0.55$ and peaks at $w_R = 0.70$; heavier XGBoost weighting ($w_R < 0.50$) degrades performance, consistent with its weaker standalone accuracy.

VI. DISCUSSION

A. Implications for Academic Integrity

Within the bounds of our synthetic setting, multi-class contamination detection appears feasible with current NLP techniques: accuracy approaching 80% across seven levels, well above the 14.3% random baseline, is potentially actionable. The strong binary performance (F1 = 0.876) suggests a practical two-stage deployment: binary screening to flag submissions, then multi-class estimation to inform a proportional response, aligning with the graduated policy frameworks advocated by Cotton et al. [3] and Perkins et al. [17].

B. The Importance of Semantic Features

The poor performance of surface-level linguistic features (39.4% accuracy) contrasts sharply with RoBERTa’s semantic analysis (79.8%). The practical implication is that simple feature-based detectors, though cheap and interpretable, cannot reliably determine contamination levels: effective detection requires deep contextual models. We note, however, that under our construction such “semantic” signal is partly entangled with topic discontinuity between blended sources, as discussed in the limitations below.

C. Limitations

Several limitations should be acknowledged, and they bound the claims above. **Construct validity.** Because each hybrid essay blends a randomly paired human and AI essay, almost always on different topics, intermediate-contamination texts contain topic and style discontinuities at sentence seams. Part

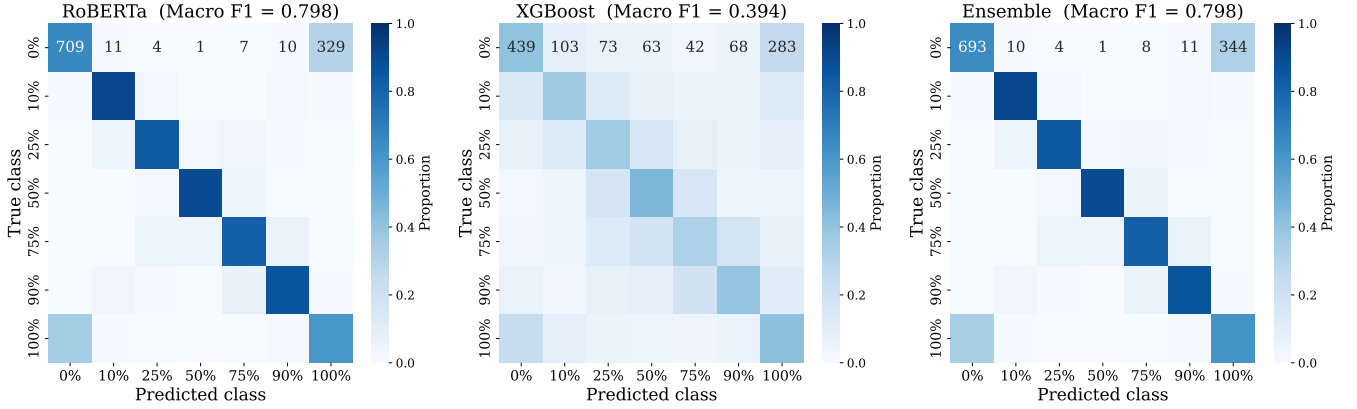


Fig. 1. Confusion matrices for all three models. Cell values show absolute counts; shading indicates normalized proportions. The RoBERTa and ensemble models exhibit strong diagonal dominance for intermediate classes (10–90%), with primary confusion between the 0% and 100% boundary classes.

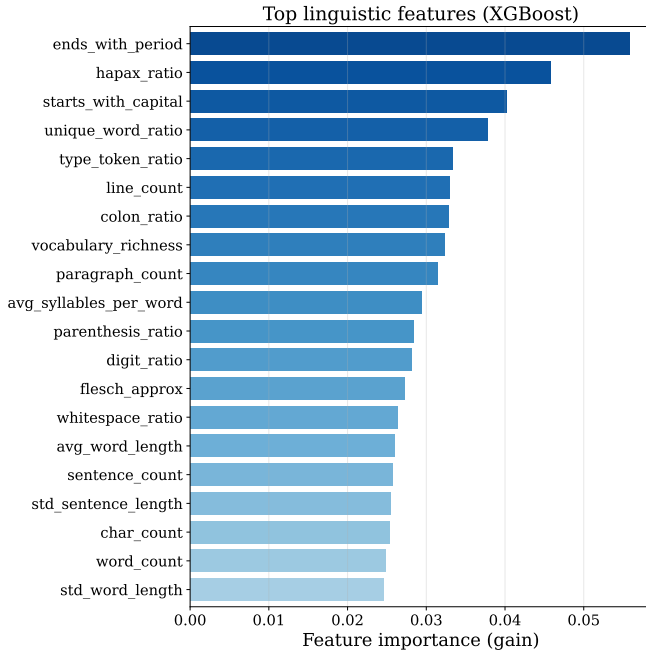


Fig. 2. Top linguistic features ranked by XGBoost importance (gain). Vocabulary richness and structural features dominate over explicit AI markers.

of the model’s discriminative power may therefore reflect the detection of internal heterogeneity rather than AI authorship as such. The pattern in Fig. 1 is consistent with this reading: the homogeneous 0% and 100% classes, which lack seams, are confused with each other, while the heterogeneous intermediate classes are easy. The reported $\sim 80\%$ accuracy should thus be interpreted as performance on this controlled construction, not as a validated estimate of AI proportion in coherent text. **External validity.** The model is trained, validated, and tested on a single synthetic generator; we do not evaluate it on authentic hybrid student essays or out-of-distribution data, and we do not run commercial binary detectors (e.g., GPTZero, Turnitin) on our data, so our comparison to them remains

conceptual rather than empirical. **Possible leakage.** We split the generated hybrids rather than the underlying source essays, so sentences from a given source essay may appear in both training and test partitions; although every hybrid is unique, this could modestly inflate performance, and source-level splitting is preferable in future work. **Dataset provenance.** The AI portion of the source corpus was produced by earlier-generation models, whereas we motivate the work with current systems (ChatGPT, Claude, Gemini); the surface traces the model learns may not transfer to contemporary generations, requiring periodic retraining. **Task formulation.** We treat contamination as nominal multi-class and optimize macro F1, which penalizes adjacent and distant level errors equally; an ordinal formulation (e.g., ordinal regression) that respects the natural ordering of levels is a promising refinement, partially motivated by the MAE results in Section V. Finally, the evaluation is restricted to English-language essays, and truncation to 512 tokens means contamination concentrated in the later parts of long essays may be missed.

D. Future Work

We identify several directions for future research, which together would address the validity limitations above and form the basis of a journal-length study. Same-topic and paragraph-level blending, paraphrase-based contamination, and human-edited AI text would produce more realistic hybrids and disentangle topic-discontinuity cues from genuine AI signal. Validation on authentic student essays and out-of-distribution corpora, together with an empirical head-to-head comparison against commercial binary detectors, would establish real-world utility. Recasting the problem as ordinal estimation, and detecting specific LLM sources within hybrid texts, are further avenues. Finally, longitudinal studies tracking robustness as LLMs evolve would inform deployment sustainability.

VII. CONCLUSION

We presented HybridDetect, a multi-class classification framework for detecting varying levels of AI contamination in

student essays. Moving beyond binary AI detection, our approach estimates contamination across seven levels (0–100%), enabling proportional institutional responses. The ensemble of RoBERTa and XGBoost achieves 79.9% accuracy and 0.798 macro F1 on 7-class classification, with 0.876 binary F1 for distinguishing predominantly human from predominantly AI text. Our analysis reveals that semantic features are essential for fine-grained contamination detection, while surface-level linguistic features provide limited discriminative power. These findings support the development of nuanced academic integrity tools that distinguish between levels of AI assistance, toward more equitable and proportional detection.

DATA AND CODE AVAILABILITY

To support reproducibility, the hybrid-essay generation pipeline, the trained RoBERTa and XGBoost models, and the evaluation scripts are released at <https://github.com/DanielDanLab/HybridDetect>. The underlying source corpus is publicly available on Kaggle [19]; our pipeline regenerates the hybrid dataset from it deterministically with a fixed random seed.

ACKNOWLEDGMENTS

This research was partially supported by the EU under Horizon Europe, grant no. 101086321 OMINO. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU or the European Research Executive Agency.

The authors used Anthropic’s Claude Sonnet 4.6 (a generative AI tool) for literature exploration, code development, manuscript drafting, and language polishing. All conceptualization, methodology, experimental validation, interpretation, and final content were reviewed, verified, and approved by the authors, who bear full responsibility for the work.

REFERENCES

- [1] E. Tian, “GPTZero: An AI text detector,” <https://gptzero.me>, 2023, accessed: 2026-03-01.
- [2] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2301.11305 pp. 24 950–24 962.
- [3] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,” *Innovations in Education and Teaching International*, vol. 61, no. 2, pp. 228–239, 2024. doi: 10.1080/14703297.2023.2190148
- [4] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2301.10226 pp. 17 061–17 084.
- [5] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, “GPT detectors are biased against non-native English writers,” *Patterns*, vol. 4, no. 7, p. 100779, 2023. doi: 10.1016/j.patter.2023.100779
- [6] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-generated text be reliably detected?” *arXiv preprint arXiv:2303.11156*, 2023. doi: 10.48550/arXiv.2303.11156
- [7] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, and L. Waddington, “Testing of detection tools for AI-generated text,” *International Journal for Educational Integrity*, vol. 19, no. 26, 2023. doi: 10.1007/s40979-023-00146-z
- [8] P. Gryka, K. Gradoń, M. Kozłowski, M. Kutyla, and A. Janicki, “Impact of spelling and editing correctness on detection of LLM-generated emails,” in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 39, 2024. doi: 10.15439/2024F8906 pp. 603–608.
- [9] Turnitin, “AI writing detection,” <https://www.turnitin.com/solutions/ai-writing>, 2023, accessed: 2026-04-01.
- [10] Winston AI, “AI content detection for education,” <https://gowinston.ai>, 2024, accessed: 2026-04-01.
- [11] Copyleaks, “AI content detector,” <https://copyleaks.com/ai-content-detector>, 2024, accessed: 2026-04-01.
- [12] L. Dugan, D. Ippolito, A. Kirubakaran, S. Shi, and C. Callison-Burch, “Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023. doi: 10.1609/aaai.v37i11.26501 pp. 12 763–12 771.
- [13] J. W. Cutler, L. Dugan, S. Havaldar, and A. Stein, “Automatic detection of hybrid human-machine text boundaries,” University of Pennsylvania, CIS 520 Course Project, 2021, available: https://www.seas.upenn.edu/~steinad/papers/cis520_project.pdf.
- [14] A. Uchendu, T. Le, and D. Lee, “Attribution and obfuscation of neural text authorship: A data mining perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 1, pp. 1–18, 2023. doi: 10.1145/3606274.3606276
- [15] D. Baidoo-Anu and L. O. Ansah, “Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning,” *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023. doi: 10.61969/jai.1337500
- [16] A. Smolansky, A. Cram, C. Radulescu, S. Zeivots, E. Huber, and R. F. Kizilcec, “Educator and student perspectives on the impact of generative AI on assessments in higher education,” in *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S)*, 2023. doi: 10.1145/3573051.3596191 pp. 378–382.
- [17] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, “The artificial intelligence assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment,” *Journal of University Teaching and Learning Practice*, vol. 21, no. 6, 2024. doi: 10.53761/q3azde36
- [18] D. Dan, A. Wróblewska, B. Grabek, M. Taczala, and M. Nakayama, “Applications and challenges of artificial intelligence in educational course design and delivery,” in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 43, 2025. doi: 10.15439/2025F8073 pp. 675–680.
- [19] M. Ejaz, “AI vs human text dataset,” Kaggle, 2024, <https://www.kaggle.com/datasets/muqaddasejaz/ai-vs-human-text-dataset>.
- [20] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, “How close is ChatGPT to human experts? comparison corpus, evaluation, and detection,” *arXiv preprint arXiv:2301.07597*, 2023. doi: 10.48550/arXiv.2301.07597
- [21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019. doi: 10.48550/arXiv.1907.11692
- [22] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785 pp. 785–794.
- [23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020. doi: 10.18653/v1/2020.emnlp-demos.6 pp. 38–45.