Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmark Evaluation of the WordBinary AI-Text Detection Model on 9,000 AI-Generated Texts and 2,000 Pre-2015 Human Academic-Paper Samples

Kiah Curan

kiah.curan@gmail.com

Independent Researcher

Research Article

Keywords: AI-generated text detection, machine-generated text, academic integrity, artificial intelligence detection, false-positive rate, humanised AI, authorship analysis, WordBinary, large language models

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10448398/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The growing use of large language models has increased demand for systems capable of distinguishing AI-generated text from genuine human writing. Such systems must achieve high sensitivity to machine-generated content while maintaining a sufficiently low false-positive rate to protect human authors. This study evaluated the WordBinary AI-text detection model using 11,000 labelled text samples: 9,000 AI-generated samples and 2,000 samples obtained from pre-2015 human academic papers. The AI corpus included normal and humanised outputs associated with Claude, Cohere, DeepSeek, Gemini, Mistral, OpenAI and OpenRouter platform groups. The human corpus contained 400 samples from each of five academic disciplines: computer science, economics, statistics, physics and mathematics.

The model classified all 9,000 AI-generated samples as AI and all 2,000 human academic samples as human. The resulting confusion matrix contained 9,000 true positives, 2,000 true negatives, no false negatives and no false positives. Observed sensitivity, specificity, precision, negative predictive value, F1 score, balanced accuracy and overall accuracy were therefore all 100%. Exact 95% binomial confidence intervals were 99.959% to 100% for sensitivity, 99.816% to 100% for specificity and 99.966% to 100% for overall accuracy. The one-sided implication of observing no false positives in 2,000 human samples is that the 95% upper confidence bound for the underlying false-positive rate is approximately 0.184%, subject to the assumption that observations were statistically independent.

Platform-level mean AI scores ranged from 98.56% for DeepSeek-associated outputs to 99.98% for OpenAI-associated outputs. All normal and humanised AI groups achieved 100% observed recall. The highest AI score among the human academic samples was 1.07%, while approximate domain-level mean scores ranged from 0.01% to 0.31%. The results demonstrate complete classification of the supplied test set. They do not, however, establish universal 100% accuracy. Independent external validation, documentation of the decision threshold, control for training-test contamination, clustered statistical analysis, multilingual testing and evaluation on mixed-authorship and AI-polished documents are required before broader generalisation.

1. Introduction

Generative artificial intelligence systems can now produce essays, summaries, reports, research-style explanations, technical material and professional communications with substantial fluency. The increasing quality and accessibility of such systems have created legitimate concerns about undeclared AI assistance, academic integrity, publication provenance, misinformation and the reliability of authorship claims.

AI-text detection is nevertheless a difficult classification problem. Human and machine writing are not fixed or mutually uniform categories. Human writing varies by discipline, educational background, language proficiency, purpose, editing history and genre. Machine-generated writing varies by model architecture, prompt, decoding configuration, system instructions, temperature, post-processing and the

extent of human revision. A detector that performs well on one generator, genre or prompting strategy may not generalise to another.

Contemporary detection methods use several technical approaches. DetectGPT, for example, examines the curvature of a language model's probability function around a passage. Ghostbuster combines features derived from multiple weaker language models and uses a trained classifier, while Binoculars contrasts the behaviour of two related language models without requiring model-specific training data. These approaches demonstrate that machine-generated text can often be detected, but they also illustrate that "AI detection" is not a single standardised method.

High classification accuracy in an internal or in-domain dataset does not necessarily imply equivalent performance under real-world distribution shifts. The RAID benchmark found that detectors could be affected by unseen generators, adversarial modifications, decoding strategies and other perturbations. M4GT-Bench similarly showed that performance commonly depends on whether the testing data resemble the domains and generators represented during training. Practical evaluations have also found substantial deterioration under modest evasion attempts and have emphasised reporting sensitivity at controlled false-positive rates rather than accuracy alone.

False positives are especially consequential in education and research. Incorrectly identifying human writing as AI-generated can affect assessment, disciplinary action, reputation and trust. Previous work has reported disproportionately high false-positive rates for some non-native English writers, although other carefully sampled assessment studies have found that well-constructed detectors can achieve strong performance without equivalent bias. These differing findings show that detector fairness depends on the model, dataset, sampling process and target population rather than on a single universal property of AI detection.

Against this background, the present study evaluates WordBinary's latest tested AI-text detection model using a supplied corpus of 9,000 AI-generated samples and 2,000 pre-2015 human academic-paper samples. The evaluation considers both sides of the classification problem: whether AI-generated writing is detected and whether genuine human academic writing is protected from false classification.

The supplied evaluation records report that all 9,000 AI samples were detected and all 2,000 human samples were retained as human. They also report that the highest AI score among the human samples was 1.07%.

1.1 Terminological clarification

This paper uses the following terms:

AI-generated sample means a text labelled in the supplied dataset as having been generated through an AI platform or model.

Normal AI sample means an AI output that was evaluated without the additional humanisation process identified in the dataset.

Humanised AI sample means an AI-generated output that underwent a rewriting or humanisation procedure before evaluation. The exact humanisation procedures were not documented in the supplied records.

Human academic-paper sample means a text sample obtained from a pre-2015 academic paper. The available evidence supports the existence of 2,000 samples, but does not establish that these samples came from 2,000 unique complete papers.

AI score means the numerical AI-likelihood output recorded by the WordBinary model. The exact calibration interpretation and classification threshold must be documented before publication.

2. Study Objectives and Research Questions

The primary objective was to measure the classification performance of the evaluated WordBinary model on the supplied labelled dataset.

The study addressed five research questions:

RQ1. What proportion of the 9,000 AI-generated samples did the WordBinary model classify as AI?

RQ2. What proportion of the 2,000 pre-2015 human academic-paper samples did the model classify as human?

RQ3. Did observed recall vary across the seven AI platform groups?

RQ4. Did observed recall differ between normal AI and humanised AI samples?

RQ5. What AI scores did the model assign to human academic writing across the five represented disciplines?

The corresponding hypotheses were:

H1. The model would achieve high recall on AI-generated text across multiple AI platform groups.

H2. Humanisation would reduce mean AI scores relative to unmodified AI output, but the model would retain substantial detection capability.

H3. The false-positive rate on pre-2015 academic writing would remain low across all five disciplines.

3. Materials and Methods

3.1 Study design

This was a retrospective benchmark evaluation of a binary text-classification system. Each sample had a reference label of either AI-generated or human-written. The WordBinary model generated a classification and an AI score, after which the classification was compared with the reference label.

The study evaluated performance at four levels:

1. Entire-dataset performance

2. AI platform-level performance

3. Model-and-mode performance

4. Human academic-domain performance

No comparison with competing commercial or open-source detectors was included in the supplied evaluation. The study therefore measures the internal performance of the tested WordBinary model and is not a comparative detector benchmark.

3.2 Dataset composition

The complete dataset contained 11,000 samples.

Reference class	Number of samples	Share of dataset
AI-generated	9,000	81.82%
Human academic	2,000	18.18%
Total	**11,000**	**100.00%**

The original evaluation describes the dataset as 9,000 AI-generated texts and 2,000 pre-2015 academic-paper samples. The human material covered computer science, economics, statistics, physics and mathematics.

Because the class distribution was not balanced, overall accuracy was reported alongside sensitivity, specificity, precision, negative predictive value, balanced accuracy and the Matthews correlation coefficient.

3.3 AI-generated corpus

The AI-generated corpus contained outputs associated with seven platform categories.

Platform group	Samples	Percentage of AI corpus
Claude	1,600	17.78%
Cohere	1,600	17.78%
DeepSeek	1,600	17.78%
Gemini	700	7.78%
Mistral	1,600	17.78%
OpenAI	1,600	17.78%
OpenRouter	300	3.33%
Total	**9,000**	**100.00%**

The corpus contained both normal AI and humanised AI material. According to the detailed aggregate table, 4,600 samples were humanised and 4,400 were normal AI outputs.

The exact prompts, system instructions, decoding parameters, temperature settings, maximum output lengths and generation dates were not included in the supplied records. The model names should consequently be treated as dataset labels until verified against generation logs and provider records.

3.4 Human academic corpus

The human reference corpus contained 2,000 pre-2015 academic-paper samples, divided evenly across five disciplines.

Academic domain	Samples	Percentage of human corpus
Computer science	400	20%
Economics	400	20%
Statistics	400	20%
Physics	400	20%
Mathematics	400	20%
Total	**2,000**	**100%**

Pre-2015 publication was used as an operational provenance criterion because the material predates the widespread public availability of modern large language model writing systems. However, publication date alone is not a complete authorship-verification method. Final reporting should document the repositories, licences, publication metadata, sampling procedure, deduplication process and any exclusion of abstracts, references, quotations, equations or machine-translated material.

3.5 Unit of analysis

The unit of analysis was one text sample. Sample lengths were not reported. This matters because AI-text detection performance frequently varies with text length, and multiple samples taken from the same source document may be statistically correlated.

If several extracts were obtained from one paper or several AI texts were produced from the same underlying prompt, the nominal sample size of 11,000 would overstate the effective number of independent observations. Future analysis should therefore retain source-paper identifiers, prompt identifiers and generation-session identifiers so that confidence intervals can be adjusted for clustering.

3.6 WordBinary model and decision rule

Each sample was processed by the WordBinary AI-text detection model. The model produced:

1. A binary classification of AI or human

2. An AI score

3. An “AI text percentage” for model-and-mode groups in the detailed records

The supplied material did not specify:

• the exact model version or software build;

• the classification threshold;

• whether the threshold was fixed before testing;

• whether the threshold had been tuned on any part of the test corpus;

• the preprocessing and text-normalisation procedure;

• minimum or maximum sample length;

• how tables, formulas, references and quotations were treated;

• whether the model had encountered any test material during training;

• whether scoring was performed in one batch or across multiple system versions; or

• whether evaluators were blinded to the reference labels.

These items are essential for replication and must be documented in the final submitted version.

3.7 Outcome definitions

For the statistical analysis, AI-generated text was treated as the positive class.

• **True positive (TP)**: AI-generated sample classified as AI

• **False negative (FN)**: AI-generated sample classified as human

• **True negative (TN)**: human sample classified as human

• **False positive (FP)**: human sample classified as AI

3.8 Evaluation measures

The following measures were calculated:

$$\text{Sensitivity or Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision or Positive Predictive Value} = \frac{TP}{TP + FP}$$

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN+FP + FN}$$

$$\text{F1} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity}+\text{Specificity}}{2}$$

The Matthews correlation coefficient was calculated as:

$$MCC = \frac{TP \times TN-FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC ranges from − 1 to 1, with 1 indicating perfect classification, 0 indicating chance-level association and − 1 indicating complete inverse classification.

3.9 Statistical analysis

Observed proportions were accompanied by two-sided 95% exact Clopper-Pearson binomial confidence intervals. Exact intervals were selected because no classification errors occurred and normal approximations can be misleading at the boundary of the probability scale.

These intervals assume statistically independent observations. They do not account for possible clustering by source paper, shared prompt, generation session, model family or humanisation method.

The supplied aggregate data were sufficient to calculate confusion-matrix measures and weighted group means. They were not sufficient to calculate:

• receiver operating characteristic curves;

• area under the ROC curve;

• precision-recall curves;

• threshold-specific sensitivity at a predetermined false-positive rate;

• calibration curves;

• Brier score;

• expected calibration error;

• score quantiles;

• standard deviations;

• confidence intervals for mean AI scores; or

• statistical significance of score differences.

These analyses require sample-level scores.

4. Results

4.1 Overall confusion matrix

The model correctly classified every sample in the supplied dataset.

Reference label	Predicted AI	Predicted human	Total
AI-generated	9,000	0	9,000
Human	0	2,000	2,000
Total	**9,000**	**2,000**	**11,000**
The resulting counts were:			
• TP = 9,000			
• FN = 0			
• TN = 2,000			
• FP = 0			

These results reproduce the supplied aggregate findings that all AI-generated texts were detected and all human academic samples were classified as human.

4.2 Overall classification measures

Measure	Observed value	95% exact confidence interval
Sensitivity/AI recall	100.000%	99.959%–100.000%
Specificity	100.000%	99.816%–100.000%
Precision	100.000%	99.959%–100.000%
Negative predictive value	100.000%	99.816%–100.000%
Overall accuracy	100.000%	99.966%–100.000%
Balanced accuracy	100.000%	Not separately estimated
F1 score	1.000	Not separately estimated
Matthews correlation coefficient	1.000	Not separately estimated
Observed false-positive rate	0.000%	0.000%–0.184%
Observed false-negative rate	0.000%	0.000%–0.041%

The 100% values are observed point estimates for this dataset. The confidence intervals demonstrate that error-free classification of a finite sample does not prove a universal error rate of zero. Under the independence assumption, the observed human result is compatible with an underlying false-positive rate as high as approximately 0.184% at the upper boundary of the two-sided 95% interval.

4.3 Platform-level AI detection

Platform group	Samples	Detected as AI	Missed as human	Recall	Mean AI score
Claude	1,600	1,600	0	100%	99.48%
Cohere	1,600	1,600	0	100%	99.94%
DeepSeek	1,600	1,600	0	100%	98.56%
Gemini	700	700	0	100%	99.89%
Mistral	1,600	1,600	0	100%	99.41%
OpenAI	1,600	1,600	0	100%	99.98%
OpenRouter	300	300	0	100%	99.30%
Total/weighted mean	**9,000**	**9,000**	**0**	**100%**	**99.50%**

All seven platform groups achieved complete observed recall. The mean platform scores ranged from 98.56% to 99.98%. The weighted mean AI score across the seven groups was approximately 99.50%.

The supplied results show that OpenAI-associated samples had the highest mean score and DeepSeek-associated samples the lowest. Even the lowest platform-level mean remained above 98%.

The platform labels do not by themselves establish out-of-distribution generalisation. That interpretation depends on whether the tested WordBinary model was trained on outputs from the same providers, model families, prompts or humanisation procedures.

4.4 Normal and humanised AI results

AI mode	Samples	Detected as AI	Missed	Recall	Weighted mean AI score	Weighted mean AI-text percentage
Humanised AI	4,600	4,600	0	100%	99.03%	86.02%
Normal AI	4,400	4,400	0	100%	99.99%	92.95%
Total	**9,000**	**9,000**	**0**	**100%**	**99.50%**	**89.41%**

The mean AI score for humanised samples was approximately 0.96 percentage points lower than the mean for normal AI samples. The mean AI-text percentage was approximately 6.93 percentage points lower for humanised samples.

Despite these score differences, every sample in both modes was classified as AI. The complete model-and-mode results in the supplied dataset report 100% recall for each listed subgroup.

A formal test of the difference between normal and humanised scores could not be conducted from group means alone. Such a comparison would require individual scores, variances and information about how samples were paired or generated.

4.5 Detailed model-and-mode results

Platform	Dataset model label	Mode	Samples	Recall	Mean AI score	Mean AI-text percentage
Claude	claude-fable-5	Humanised	200	100%	96.74%	72.59%
Claude	claude-fable-5	Normal	200	100%	100.00%	89.66%
Claude	claude-haiku-4-5-20251001	Humanised	200	100%	99.76%	94.38%
Claude	claude-haiku-4-5-20251001	Normal	200	100%	99.92%	84.54%
Claude	claude-sonnet-4-5-20250929	Humanised	200	100%	99.60%	89.22%
Claude	claude-sonnet-4-5-20250929	Normal	200	100%	100.00%	99.21%
Claude	claude-sonnet-5	Humanised	200	100%	99.78%	84.08%
Claude	claude-sonnet-5	Normal	200	100%	100.00%	99.34%
Cohere	command-a-03-2025	Humanised	400	100%	99.88%	88.43%
Cohere	command-a-03-2025	Normal	400	100%	100.00%	92.28%
Cohere	command-r-08-2024	Humanised	400	100%	99.90%	79.94%
Cohere	command-r-08-2024	Normal	400	100%	100.00%	93.10%
DeepSeek	deepseek-chat	Humanised	400	100%	99.27%	85.99%
DeepSeek	deepseek-chat	Normal	400	100%	100.00%	87.91%
DeepSeek	deepseek-reasoner	Humanised	400	100%	94.98%	78.83%
DeepSeek	deepseek-reasoner	Normal	400	100%	99.99%	87.38%
Gemini	gemini-2.5-flash	Humanised	200	100%	100.00%	83.69%
Gemini	gemini-2.5-flash-lite	Normal	100	100%	100.00%	100.00%
Gemini	gemini-flash-lite-latest	Humanised	200	100%	99.62%	89.92%
Gemini	gemini-flash-lite-latest	Normal	200	100%	100.00%	99.12%
Mistral	magistral-small-latest	Humanised	200	100%	99.66%	83.59%

Platform	Dataset model label	Mode	Samples	Recall	Mean AI score	Mean AI-text percentage
Mistral	magistral-small-latest	Normal	200	100%	100.00%	95.78%
Mistral	ministral-8b-latest	Humanised	200	100%	98.73%	90.12%
Mistral	ministral-8b-latest	Normal	200	100%	100.00%	96.20%
Mistral	mistral-medium-latest	Humanised	200	100%	97.54%	81.41%
Mistral	mistral-medium-latest	Normal	200	100%	100.00%	97.15%
Mistral	mistral-small-latest	Humanised	200	100%	99.36%	78.65%
Mistral	mistral-small-latest	Normal	200	100%	100.00%	96.38%
OpenAI	gpt-4.1-mini	Humanised	200	100%	100.00%	94.55%
OpenAI	gpt-4.1-mini	Normal	200	100%	100.00%	86.94%
OpenAI	gpt-4.1-nano	Humanised	200	100%	100.00%	95.08%
OpenAI	gpt-4.1-nano	Normal	200	100%	100.00%	99.34%
OpenAI	gpt-4o-mini	Humanised	200	100%	99.91%	85.11%
OpenAI	gpt-4o-mini	Normal	200	100%	100.00%	94.78%
OpenAI	gpt-5-mini	Humanised	200	100%	99.98%	93.33%
OpenAI	gpt-5-mini	Normal	200	100%	99.97%	86.06%
OpenRouter	google/gemma-4-26b-a4b-it	Humanised	200	100%	98.95%	96.25%
OpenRouter	google/gemma-4-31b-it	Normal	100	100%	100.00%	98.04%

The lowest reported subgroup mean AI score was 94.98% for the humanised DeepSeek Reasoner group. Several groups recorded mean scores of 100.00%. These are subgroup means and must not be interpreted as minimum or maximum individual scores.

4.6 Human academic-domain results

Human academic domain	Samples	Correctly classified as human	False positives	Mean AI score
Computer science	400	400	0	Approximately 0.05%
Economics	400	400	0	Approximately 0.31%
Statistics	400	400	0	Approximately 0.06%
Physics	400	400	0	Approximately 0.06%
Mathematics	400	400	0	Approximately 0.01%
Total/weighted mean	**2,000**	**2,000**	**0**	**Approximately 0.10%**

All human academic samples were classified as human. The highest individual AI score in the human corpus was reported as 1.07%. Economics had the highest approximate domain mean at 0.31%, while mathematics had the lowest at approximately 0.01%.

The absence of false positives across these five fields indicates strong protection of the represented human academic material. It does not establish equivalent performance for student essays, reflective writing, creative writing, correspondence, translated material, second-language writing or other academic disciplines.

4.7 Separation of reported scores

The AI platform groups had mean AI scores between 98.56% and 99.98%. The human academic domains had approximate mean scores between 0.01% and 0.31%, and the highest reported human score was 1.07%.

This is substantial aggregate score separation. However, complete distributional separation cannot be formally quantified without the individual AI scores. The minimum AI-generated score was not reported, and platform-level means cannot be substituted for individual minima.

5. Discussion

5.1 Principal findings

The evaluated WordBinary model classified all 11,000 supplied samples correctly. It detected all 9,000 AI-generated samples and retained all 2,000 pre-2015 human academic-paper samples as human. Observed sensitivity, specificity and accuracy were therefore each 100%.

Three findings are especially notable.

First, complete observed recall was maintained across seven AI platform categories. The corpus was not restricted to one provider label or one model family. This is relevant because detectors trained narrowly on one generator can experience reduced recall when evaluated on outputs from other models.

Second, all 4,600 humanised AI samples were detected. Humanisation reduced aggregate AI-score and AI-text-percentage measures, but it did not change any binary classifications in this dataset.

Third, no human academic sample was flagged as AI. This is important because a detector can obtain high AI recall simply by applying an overly aggressive classification rule. The human results provide evidence that the observed recall was not achieved at the cost of false positives within the five represented domains.

5.2 Interpretation of the perfect point estimates

A result of 11,000 correct classifications out of 11,000 is statistically and operationally strong. It must nevertheless be described as **100% observed accuracy in the supplied dataset**, not as proof that the model is universally 100% accurate.

Even with no observed errors, the exact confidence interval for overall accuracy has a lower boundary below 100%. The specificity interval is wider than the sensitivity interval because the human set contained fewer observations. More importantly, the binomial intervals assume independent samples. If many samples were derived from common documents, prompts or transformation templates, the effective evidence would be lower than the nominal sample count suggests.

A perfect result can also arise when the test distribution closely matches the model's training distribution. Robust external evaluation therefore requires data that were excluded from training, threshold selection and model development.

5.3 Performance on humanised AI

The complete recall on humanised samples is an important finding because post-generation rewriting is a known challenge for automated detection. Benchmark research has shown that paraphrasing, humanisation and relatively modest adversarial modifications can weaken many detectors.

In the WordBinary dataset, the weighted mean AI score declined from approximately 99.99% for normal AI output to 99.03% for humanised output. The mean AI-text percentage declined more substantially, from approximately 92.95% to 86.02%. These changes suggest that the humanisation procedure modified at least some of the signals used by the model, while leaving sufficient evidence for classification.

The interpretation is limited by the absence of details about the humanisation process. "Humanised" may refer to manual rewriting, an automated paraphraser, a dedicated AI humaniser, a second LLM or a

combination of interventions. Robustness to one transformation should not be treated as robustness to all transformations.

5.4 Human-text protection

The model's human-text performance is particularly relevant to academic use. All 2,000 samples were classified as human, and the highest recorded human AI score was 1.07%. This suggests that the model did not mistake the formal, repetitive or technically structured conventions of the five disciplines for machine authorship.

The result is encouraging but narrower than a full fairness evaluation. Published studies have shown that detector behaviour can vary according to language proficiency, language background, author characteristics and domain. Some studies have reported significant disparities, while others have found that carefully sampled and appropriately trained systems can avoid systematic disadvantage.

The present human corpus does not appear to have been stratified by first language, country, writing proficiency, sex, race or ethnicity. No conclusions about demographic fairness should therefore be drawn from the current results.

5.5 Relationship to existing benchmarks

Recent AI-text detection research increasingly emphasises multi-generator, multi-domain and adversarial testing. RAID includes millions of generations across models, domains, attacks and decoding configurations. M4GT-Bench evaluates multilingual, multi-domain, multi-generator and mixed human-machine detection. These benchmarks show that apparently strong detectors can deteriorate when test conditions differ from development data.

The current WordBinary evaluation provides substantial sample volume and useful variation across platform and mode labels. It is strongest as an internal benchmark of the tested corpus. It does not yet provide the methodological controls needed to claim equivalent performance on RAID, M4GT-Bench or an independently administered out-of-distribution benchmark.

5.6 AI-polished human writing

The current binary design separates fully AI-generated material from human academic material. It does not address the increasingly common category of AI-polished human writing.

A human author may write an original document and use an AI system only to correct grammar, improve clarity, shorten sentences or adjust tone. Recent research indicates that detectors can struggle to distinguish minimally AI-polished human text from substantially generated material. This creates both a technical and policy problem: the detector's label may not correspond neatly to the institution's definition of permitted or prohibited assistance.

Future WordBinary studies should treat at least four classes separately:

1. Fully human-written text

2. Human-written text corrected only for spelling or grammar

3. Human-written text substantially rewritten or polished by AI

4. Fully or predominantly AI-generated text

5.7 Practical meaning of an AI-detection result

An AI score should be interpreted as a model-generated review signal rather than direct proof of misconduct or authorship. A detector ordinarily evaluates linguistic or statistical characteristics of the submitted text. It does not observe the writing process, establish who entered the words, determine whether AI use was authorised or assess the writer's intent.

In academic-integrity settings, a detection result should be considered alongside:

• drafts and version history;

• notes and research records;

• source use and citation behaviour;

• oral explanation of the work;

• consistency with previous writing;

• institutional rules concerning permitted AI assistance; and

• contextual human review.

The present findings support the model's potential utility as part of such a workflow. They do not justify using the score as the sole basis for an adverse academic or professional decision.

6. Limitations

6.1 Internal rather than independent evaluation

The records describe testing of WordBinary's model on a WordBinary evaluation dataset. No independent institution, external laboratory or blinded third party is identified. Independent replication is necessary to reduce the risk of unintentional selection, implementation or reporting bias.

6.2 Unreported model version

The phrase "latest model" is not reproducible. AI detection systems may change after retraining, threshold adjustment, preprocessing changes or software deployment. The final paper must state the

exact model version, release identifier, code commit, configuration checksum and evaluation date.

6.3 Unknown relationship to training data

The supplied material does not establish that every test sample was excluded from model training, feature engineering, threshold selection and error analysis. A strict locked holdout dataset should be created before testing.

Where the detector was developed using outputs from the same generator families, the evaluation should distinguish:

• seen generator and seen domain;

• seen generator and unseen domain;

• unseen generator and seen domain;

• unseen generator and unseen domain.

6.4 Potential non-independence

The number of unique source papers, prompts, topics and generation sessions was not reported. Multiple extracts from one paper or multiple generations from one prompt can be highly correlated. Treating correlated samples as independent narrows confidence intervals and may exaggerate the apparent diversity of the test.

6.5 Human provenance

Pre-2015 publication strongly reduces the likelihood of modern LLM-generated authorship, but publication date does not independently prove that every extracted passage was written entirely by a human. The human corpus requires auditable bibliographic provenance and source-level deduplication.

6.6 Sample length

Text lengths and length distributions were not supplied. Detection performance often varies materially between short and long passages. Results should be stratified into predefined word-count ranges.

6.7 Humanisation procedure

The humanisation methods were not described. The results therefore establish performance only for the unspecified transformations represented in the supplied corpus.

6.8 English-language and domain scope

The supplied dataset appears to evaluate academic English. It does not establish multilingual performance. The human corpus covers five quantitative or technical disciplines but does not include law, medicine, education, management, psychology, sociology, history, literature or other humanities and social-science fields.

6.9 Mixed-authorship documents

The test set used binary samples labelled as AI or human. Real documents may contain varying proportions of AI-generated and human-written content. The results do not establish sentence-level localisation or reliable estimation of the proportion of AI-generated content in a mixed document.

6.10 Score calibration

The evaluation reported high AI scores for AI text and low scores for human text, but calibration was not assessed. A numerical value of 90%, for example, should not automatically be interpreted as a literal 90% probability of AI authorship unless calibration has been empirically demonstrated.

6.11 No comparator systems

The study did not evaluate other detectors on the same samples. It therefore cannot conclude that WordBinary outperforms alternative commercial or open-source systems.

6.12 No sample-level uncertainty analysis

Only aggregate means and classification counts were available. Standard deviations, medians, interquartile ranges, individual minima and maxima, calibration measures and threshold curves could not be calculated.

These limitations are consistent with the scope restrictions identified in the supplied study material, including the absence of mixed-authorship, multilingual and broader disciplinary testing.

7. Recommended External Validation Protocol

7. Recommended External Validation Protocol
A subsequent preregistered evaluation should use a frozen WordBinary model and a dataset assembled after the model has been locked.

7.1 Dataset structure

A recommended dataset would contain:

Evaluation category	Recommended minimum
Fully human texts	5,000
Fully AI-generated texts	5,000
AI-humanised texts	2,000
AI-polished human texts	2,000
Mixed-authorship texts	2,000
Translated texts	1,000
Non-native English texts	2,000
Multilingual texts	At least 1,000 per language

7.2 Generator coverage

The AI corpus should include:

• generators included during model development;

• generators entirely unseen during training;

• open-weight and closed-source systems;

• multiple model sizes;

• multiple decoding settings;

• multiple prompt structures;

• factual, argumentative, reflective and creative tasks; and

• outputs produced at different dates.

7.3 Humanisation conditions

Humanisation should be divided into reproducible categories:

• manual light editing;

• manual substantial rewriting;

• automated paraphrasing;

• AI-based rewriting;

• dedicated humaniser tools;

• translation and back-translation;

• sentence reordering;

• synonym substitution; and

• combined attacks.

7.4 Human-text coverage

Human material should include:

• native and non-native English writers;

• multiple proficiency levels;

• student and professional writing;

• regional English varieties;

• multiple academic and non-academic domains;

• edited and unedited writing;

• spelling- and grammar-corrected writing; and

• writing produced under controlled conditions.

7.5 Preregistration and blinding

Before evaluation, the following should be frozen and timestamped:

• detector version;

• threshold;

• preprocessing;

• exclusion rules;

• primary outcome;

• sample-size calculation;

• subgroup definitions; and

• statistical analysis plan.

The person administering the detector should not know the reference labels, and the reference-label adjudicators should not see WordBinary scores.

7.6 Reporting measures

Future reports should present:

• sensitivity;

• specificity;

• precision;

• negative predictive value;

• F1 score;

• Matthews correlation coefficient;

• AUROC;

• area under the precision-recall curve;

• TPR at predefined FPR levels;

• calibration;

• bootstrap or cluster-robust confidence intervals;

• subgroup fairness measures;

• length-stratified results; and

• performance on generators and domains excluded from training.

False-positive-controlled reporting is particularly important because a small false-positive rate can affect many people when a detector is used at institutional scale. Recent research has proposed methods specifically designed to constrain false-positive behaviour rather than optimise unconstrained accuracy alone.

8. Practical and Ethical Implications

8.1 Academic institutions

The result supports further evaluation of WordBinary as a component of academic-integrity review. Institutions should define in advance:

• what AI assistance is permitted;

• what score initiates review;

• what corroborating evidence is required;

• how students may respond;

• who makes the final determination; and

• how detector limitations are disclosed.

8.2 Students and authors

Writers may use the model as a diagnostic aid to identify passages that resemble patterns associated with AI generation. A score should not be presented as a definitive authorship verdict.

8.3 Publishers and research organisations

Publishers may use AI detection to prioritise manual checks, but the model should not replace plagiarism analysis, citation verification, data-integrity checks, image-forensics review or editorial judgement.

8.4 Transparency

Reports should clearly distinguish:

• document-level AI score;

• proportion of text highlighted;

• sentence-level evidence;

• uncertainty;

• excluded material; and

• the difference between AI indication and proof of misconduct.

8.5 Human oversight

High-stakes determinations should remain subject to meaningful human review. Automated classification can inform an inquiry, but disciplinary or reputational decisions require contextual evidence and procedural fairness.

9. Conclusion

This study evaluated the WordBinary AI-text detection model on 11,000 labelled samples: 9,000 AI-generated texts and 2,000 pre-2015 human academic-paper samples.

The model classified all 9,000 AI samples as AI and all 2,000 human samples as human. The resulting observed sensitivity, specificity, precision, F1 score, balanced accuracy and overall accuracy were 100%, with a Matthews correlation coefficient of 1.000. The exact 95% confidence interval for sensitivity was 99.959% to 100%, while the interval for specificity was 99.816% to 100%.

Complete recall was observed across Claude, Cohere, DeepSeek, Gemini, Mistral, OpenAI and OpenRouter platform groups. It was also observed across both normal and humanised AI samples. Platform-level mean AI scores ranged from 98.56% to 99.98%, and the weighted mean AI score across all AI platform groups was approximately 99.50%.

All 2,000 human academic samples were classified as human. The highest reported AI score among these samples was 1.07%, and the approximate domain means ranged from 0.01% to 0.31%.

These findings demonstrate complete classification of the supplied benchmark dataset and provide strong evidence of performance under its represented conditions. They should not be interpreted as proof of universal 100% accuracy. The study remains limited by the absence of independent administration, an exact model identifier, sample-level score data, documented generation and humanisation procedures, training-test separation evidence, clustered analysis, multilingual material, AI-polished human writing and mixed-authorship documents.

The appropriate conclusion is therefore that WordBinary achieved **100% observed classification accuracy on this specific 11,000-sample evaluation**, with promising results for multi-platform AI detection, humanised AI detection and protection of the represented human academic writing. Independent, preregistered and reproducible external validation is the necessary next step.

Declarations

Ethics statement

The study involved text samples and did not report intervention with human participants. The final submission should confirm that all human-paper material was obtained and processed in accordance with applicable copyright, database, privacy and research-governance requirements.

References

1. Chakraborty S, Bedi AS, Zhu S, An B, Manocha D, Huang F (2023) On the possibilities of AI-generated text detection. arXiv:2304.04736.
2. Dugan L, Hwang A, Trhlík F, Zhu A, Ludan JM, Xu H, Ippolito D, Callison-Burch C (2024) RAID: A shared benchmark for robust evaluation of machine-generated text detectors. Proceedings of the

62nd Annual Meeting of the Association for Computational Linguistics, 12463–12492. 10.18653/v1/2024.acl-long.674

3. Hans A, Schwarzschild A, Cherepanova V, Kazemi H, Saha A, Goldblum M, Geiping J, Goldstein T (2024) Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. Proceedings of the 41st International Conference on Machine Learning, 17519–17537
4. Jiang Y, Hao J, Fauss M, Li C (2024) Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native. Engl speakers? Computers Educ 217:105070. 10.1016/j.compedu.2024.105070
5. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J (2023) GPT detectors are biased against non-native English writers. Patterns 4(7):100779. 10.1016/j.patter.2023.100779
6. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C (2023) DetectGPT: Zero-shot machine-generated text detection using probability curvature. Proceedings of the 40th International Conference on Machine Learning, 24950–24962
7. Saha S, Feizi S (2025) Almost AI, almost human: The challenge of detecting AI-polished writing. Findings of the Association for Computational Linguistics: ACL 2025
8. Tufts B, Zhao X, Li L (2025) A practical examination of AI-generated text detectors for large language models. Findings of the Association for Computational Linguistics: NAACL 2025, 4839–4856. 10.18653/v1/2025.findings-naacl.271
9. Verma V, Fleisig E, Tomlin N, Klein D (2024) Ghostbuster: Detecting text ghostwritten by large language models. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, 1702–1717. 10.18653/v1/2024.naacl-long.95
10. Wang Y, Mansurov J, Ivanov P, Su J, Shelmanov A, Tsvigun A, Afzal OM, Mahmoud T, Puccetti G, Arnold T, Aji A, Habash N, Gurevych I, Nakov P (2024) M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 3964–3992. 10.18653/v1/2024.acl-long.218
11. Zhu X, Ren Y, Cao Y, Lin X, Fang F, Li Y (2025) Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics