GenAI Content Detection Task 3: Cross-Domain Machine-Generated Text Detection Challenge

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov,
Marianna Apidianaki, Chris Callison-Burch

COLING 2025

Penn
UNIVERSITY of PENNSYLVANIA

RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors

Liam Dugan[1], **Alyssa Hwang**[1], **Filip Trhlik**[2], **Josh Magnus Ludan**[1]
Andrew Zhu[1], **Hainiu Xu**[3], **Daphne Ippolito**[4], **Chris Callison-Burch**[1]
University of Pennsylvania[1] University College London[2]
King's College London[3] Carnegie Mellon University[4]
{ldugan, ahwang16, jludan, andrz, ccb}@seas.upenn.edu
hainiu.xu@kcl.ac.uk, filip.trhlik.21@ucl.ac.uk, daphnei@cmu.edu

Abstract

Many commercial and open-source models claim to detect machine-generated text with extremely high accuracy (99% or more). However, very few of these detectors are evaluated on shared benchmark datasets and even when they are, the datasets used for evaluation are insufficiently challenging—lacking variations in sampling strategy, adversarial attacks, and open-source generative models. In this work we present RAID: the largest and most challenging benchmark dataset for machine-generated text detection. RAID includes over 6 million generations spanning 11 models, 8 domains, 11 adversarial attacks and 4 decoding strategies. Using RAID, we evaluate the out-of-domain and adversarial robustness of 8 open- and 4 closed-source detectors and find that current detectors are easily fooled by adversarial attacks, variations in sampling strategies, repetition penalties, and unseen generative models. We release our data[1] along with a leaderboard[2]

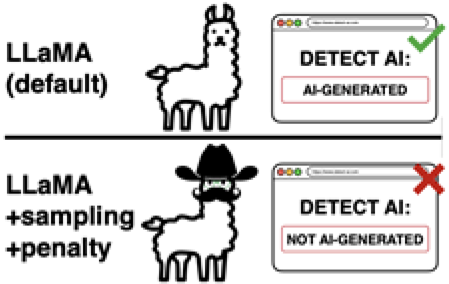


Figure 1: Detectors for machine-generated text are often highly performant on default model settings but fail to detect more unusual settings such as using random sampling with a repetition penalty.

their own evaluation datasets and fail to test their models on shared resources—making it difficult to verify claims of accuracy and robustness. This has led to an erosion of trust in the efficacy of auto-

Models

ChatGPT
Cohere
MPT-30B
Mistral 7B
GPT-2 XL
GPT-3
LLaMA 2 70B (Chat)
Cohere (Chat)
MPT-30B (Chat)
Mistral 7B (Chat)
GPT-4

11 models

Domains

Abstracts
Books
News
Poetry
Czech*
Code*
Recipes
Reddit
Reviews
Wikipedia
German*

11 domains

Decoding Strategy

Greedy `(temp. = 0)`
Sampling `(temp. = 1, p = 1)`

Repetition Penalty

With ✓ `(rep = 1.2)`
Without ✗ `(rep = 1.0)`

Detectors

Neural
RoBERTa-B (GPT-2)
RoBERTa-L (GPT-2)
RoBERTa (ChatGPT)
RADAR

Metric-Based
GLTR
Fast DetectGPT
Binoculars
LLMDet

Commercial
GPT-Zero
Originality.AI
Winston.AI
ZeroGPT

12 detectors

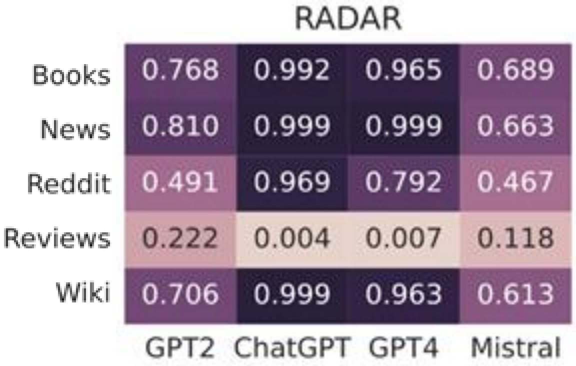
Adversarial Attacks

Alternative Spelling
Article Deletion
Insert Paragraphs
Upper Lower Swap
Zero-Width Space
Whitespace Addition
Homoglyph
Number Swap
Paraphrase
Synonym Swap
Misspelling

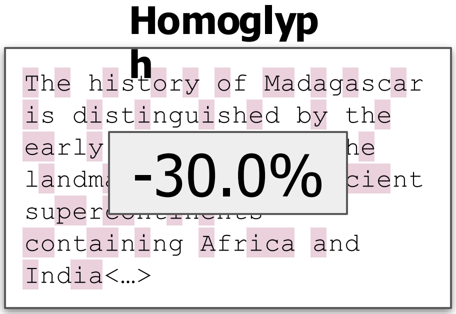
11 attacks

We found various weaknesses of classifiers

Model-Specific



Domain-Specific



Adversarial

What happens when you train on RAID?

Research Questions

1. Can a single detector be trained to detect generated text from many different known domains and LLMs accurately?

2. Can a single detector be robust to many different known adversarial attacks?

Task Setup

Phase 1: (September 18th - November 2nd) Released the training data and baseline results

Phase 2: (November 2nd - November 6th) Conduct the official evaluation on the test set and release the leaderboard

Phase 3: (November 6th - November 15th) System paper and summary paper writing stage

Subtask A

(Non-Adversarial)

- 11 LLMs
- 4 Decoding strategies
- 8 Domains

Subtask B

(Adversarial)

- 11 LLMs
- 4 Decoding strategies
- 8 Domains

\+ 12 Adversarial Attacks

Subtask A Training Data

		Human	ChatGPT	dav.-003	GPT-4	Cohere	Coh.-C	GPT-2	MPT	MPT-C	Mistral	Mist.-C	Llama2-C
Train	Abstracts	1766	3532	3532	3532	3532	3532	7064	7064	7064	7064	7064	7064
	Books	1781	3562	3562	3562	3562	3562	7124	7124	7124	7124	7124	7124
	News	1780	3560	3560	3560	3560	3560	7120	7120	7120	7120	7120	7120
	Poetry	1771	3542	3542	3542	3542	3542	7084	7084	7084	7084	7084	7084
	Recipes	1772	3544	3544	3544	3544	3544	7088	7088	7088	7088	7088	7088
	Reddit	1779	3558	3558	3558	3558	3558	7116	7116	7116	7116	7116	7116
	Wiki	1779	3558	3558	3558	3558	3558	7116	7116	7116	7116	7116	7116
	Reviews	943	1886	1886	1886	1886	1886	3772	3772	3772	3772	3772	3772
	Total	13371	26742	26742	26742	26742	26742	53484	53484	53484	53484	53484	53484

x12 to each for adversarial

Evaluation Metric

TPR @ FPR=5%

"How much AI-Generated text do you correctly detect while maintaining a 5% False Positive Rate?"

To compute:

Run detector on human-written text → Tune threshold to 5% FPR → Evaluate Recall (TPR)

Participants also got

Code for Adversarial Attacks

Source Domains for Human Text

Results

Subtask A

(Non-Adversarial)

Team Ranking (Subtask A)		
	Best Submission	**Result**
[Le] Leidos	Leidos v1.0.3	**99.4** (0.6)
[Pa] Pangram	Pangram	99.3 (**0.4**)
[Us] USTC	R-L Focal Loss	98.1 (1.3)
[Al] ALERT	ALERT v1.1	91.8 (9.4)
[Cn] CNLP	DistilBERT-NITS	90.5 (2.9)
[Lx] LuxVeri	R-B & R-Oai	82.6 (10.9)
[Ba] **Baseline**	Binoculars	79.0 (2.4)
[Mo] MOSAIC	MOSAIC-4	75.2 (5.9)
[80] 1-800	L3-60 Zero-shot	57.1 (9.6)
[Ra] Random	Adv. CDMGTD	3.2 (1.6)

Previous SoTA: 94.9

Subtask B

(Adversarial)

Team Ranking (Subtask B)		
	Best Submission	**Result**
[Le] Leidos	Leidos v1.0.2	**97.7** (**2.5**)
[Pa] Pangram	Pangram	**97.7** (2.9)
[Us] USTC	R-L Focal Loss	92.7 (9.5)
[Al] ALERT	ALERT v1.1	82.6 (15.5)
[Lx] LuxVeri	Fine-tuned R-B	80.1 (8.4)
[Ba] **Baseline**	Binoculars	71.3 (16.2)
[Mo] MOSAIC	MOSAIC-5	69.4 (16.3)
[80] 1-800	L3-60 Zero-shot	51.4 (15.4)
[Cn] CNLP	Adv.-sub.-3	41.6 (10.4)
[Ra] Random	Adv. CDMGTD	6.5 (7.6)

Previous SoTA: 86.2

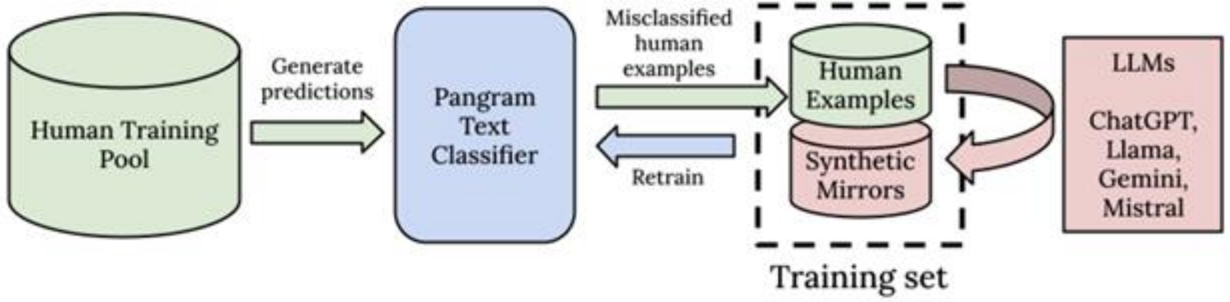
Pangram

Base Model

- Fine-tune Mistral NeMo classifier (12B params) on a large-scale corpus
  - Use LoRA training + linear classification head + LLM prediction head
- Preprocess data: (Remove zero-width, lowercase, convert unicode, etc.)

Key Insight: Hard Negative Mining

- Select the 50k examples in RAID with highest error → Re-train the model



[1] "Pangram at GenAI Detection Task 3: An Active Learning Approach to Machine-Generated Text Detection" (Emi et al. 2025)

Leidos

Base Model

- Trained four classifiers using Distil-RoBERTa-base as the base model
  - **Binary Classifier with & without Class Weighting**
  - **Multi-Class Classifier with & without Class Weighting**

Key Insight: Class Weighting

- Compute weights using the formula $w_i = N / (C \times n_i)$
  - N = total # docs, C = # classes, n_i = # docs in class i
- Upweights the loss on human text and downweights the loss on MGT

$$w_i = \frac{N}{C \times n_i}$$

[2] "Leidos at GenAI Detection Task 3: A Weight-Balanced Transformer Approach for AI Generated Text Detection Across Domains" (Edikala et al. 2025)

ALERT

Base Model

- Documents are embedded using an authorship style embedding model
  - Trained using contrastive learning on large corpus of **human authorship data**
  - Hard positive and negative mining using BM25 and k-means clustering
- Embeddings given to a single feed-forward layer and binary classification head

Key Insight: Human Authorship Data

- Style embeddings trained on only human data still separate MGT fairly well!

[3] "BBN-U.Oregon's ALERT system at GenAI Content Detection Task 3: Robust Authorship Style Representations for Cross-Domain Machine-Generated Text Detection" (Kandula et al. 2025)

Broader Trends

1. Hard Positive and Negative Sampling was broadly effective and featured in many strong teams' submissions

2. Preprocessing and normalization of input text was effective at evading simple yet powerful adversarial attacks.

3. Utilizing or creating external data was broadly helpful even when said data is from different LLMs or for a different task

4. Incredible diversity of modeling approaches!

Takeaways

(And other reflections…)

RAID was supposed to be challenging!

						Generator Model						
Detector	**Aggregate**	**chatgpt**	**gpt4**	**gpt3**	**gpt2**	**mistral**	**mistral-chat**	**cohere**	**cohere-chat**	**llama-chat**	**mpt**	**mpt-chat**
Leidos Detector v1.0.3	**0.994**	0.998	0.999	**1.000**	**0.999**	**0.995**	**1.000**	0.953	0.974	**1.000**	**0.997**	0.999
Pangram	0.993	0.998	**1.000**	0.988	0.998	0.990	0.998	**0.971**	**0.977**	**1.000**	0.992	0.998
Leidos Detector v1.0.2	0.993	**0.999**	**1.000**	**1.000**	0.995	0.985	0.999	0.966	0.967	0.998	0.995	0.999
Leidos Detector v1.0.4	0.992	0.998	**1.000**	**1.000**	0.998	0.987	0.999	0.958	0.954	0.999	**0.997**	**1.000**
Leidos Detector v1.0.1	0.991	0.998	0.999	**1.000**	0.997	0.986	**1.000**	0.942	0.965	0.999	0.995	0.998
roberta_focalloss	0.981	0.989	0.989	0.996	0.986	0.973	0.995	0.902	0.930	0.999	0.985	0.995
ALERT MGT Detector v1.1	0.918	0.976	0.943	0.917	0.919	0.862	0.973	0.706	0.848	0.988	0.905	0.960
DistilBERT-NITS	0.905	0.989	0.967	0.835	0.880	0.846	0.976	0.639	0.835	0.987	0.884	0.985
ALERT MGT Detector v1.2	0.893	0.958	0.917	0.932	0.897	0.826	0.943	0.725	0.823	0.952	0.873	0.922

We did not expect anyone to get over 99 TPR

RAID does have flaws

- Easy shortcuts for detection
  - e.g. regular \n characters, formatting errors
- Instances of Meta-Commentary
  - e.g. “Sure, I can help!”
- Degenerate / Repetitive output texts
  - Mainly for older continuation models

1. Can a single detector be trained to detect generated text from many different known domains and LLMs accurately?

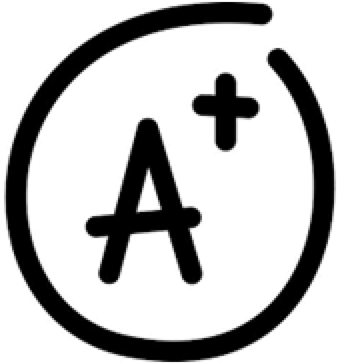
2. Can a single detector be robust to many different known adversarial attacks?

Maybe?

Building Harder Benchmarks

Filtering out text with "shortcuts"

Increasing diversity of text



More aggressive false positive rates

Thank you!

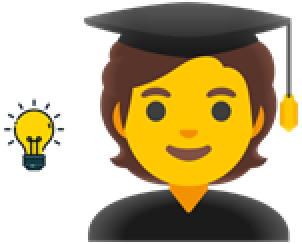
Thanks!

Shared Task

https://raid-bench.xyz/shared-task

Paper

https://arxiv.org/abs/2501.08913



Read the Paper for more!

Subtask A: Performance Across Domains (Official Results)									
	News	**Wiki**	**Reddit**	**Books**	**Abs.**	**Reviews**	**Poetry**	**Recipes**	**Total** (σ)
[Le] Leidos v1.0.3	**99.9**	99.8	98.3	99.4	99.9	98.6	99.3	**100.0**	**99.4** (0.6)
[Pa] Pangram	99.7	99.1	98.5	99.5	99.3	**99.6**	98.8	99.9	99.3 (**0.4**)
[Le] Leidos v1.0.2	**99.9**	**99.9**	**99.4**	99.5	99.9	95.9	**99.6**	**100.0**	99.3 (1.2)
[Le] Leidos v1.0.4	**99.9**	99.7	99.0	99.3	**100.0**	96.5	99.4	**100.0**	99.2 (1.1)
[Le] Leidos v1.0.1	**99.9**	99.8	98.6	99.4	99.9	96.2	99.4	**100.0**	99.1 (1.2)
[Us] R-L Focal Loss	99.0	97.8	96.1	98.1	99.8	97.0	97.0	99.9	98.1 (1.3)
[Al] ALERT v1.1	99.7	95.4	75.7	**99.9**	99.9	87.2	78.3	98.3	91.8 (9.4)
[Cn] DistilBERT-NITS	89.9	87.7	90.0	93.5	90.9	85.9	90.0	96.0	90.5 (2.9)
[Al] ALERT v1.2	99.5	91.3	87.2	99.2	99.9	89.9	64.9	82.8	89.3 (11.0)
[Lx] R-B & R-Oai	87.5	90.2	62.4	89.5	99.2	83.7	73.5	75.1	82.6 (10.9)
[Lx] R-Oai & BERT	91.8	87.3	75.1	87.1	97.0	86.0	76.3	59.4	82.5 (11.1)
[Lx] Fine-tuned R-B	87.5	89.7	61.7	89.6	98.8	82.5	66.3	74.6	81.3 (11.9)
[Ba] Binoculars	80.7	76.5	81.3	82.8	76.0	78.0	80.1	76.4	79.0 (2.4)
[Mo] MOSAIC-4	79.5	67.6	78.2	79.8	77.1	81.4	63.7	75.8	75.2 (5.9)
[Mo] MOSAIC-5	79.0	65.8	76.7	79.8	76.5	77.2	64.8	75.1	74.5 (5.4)
[Lx] Radar & R-L	91.6	73.7	76.3	78.1	74.2	58.7	45.7	73.5	71.5 (12.8)
[Ba] RADAR	87.4	77.3	73.6	78.1	67.5	6.3	46.0	88.7	65.6 (25.7)
[Ba] GLTR	67.7	63.6	63.2	71.9	60.1	65.0	18.2	67.9	59.7 (16.0)
[80] L3-60 Zero-shot	63.6	36.5	61.5	65.4	55.3	68.9	51.5	53.9	57.1 (9.6)
[80] M3-60 Zero-shot	58.1	58.1	65.8	63.3	44.1	67.1	53.2	50.5	56.5 (7.4)
[Ba] openai-roberta-large	67.8	59.4	60.0	52.5	64.8	52.8	23.3	65.1	55.7 (13.3)
[Cn] Adv.-submission-3	27.1	26.1	52.8	57.1	30.1	48.6	38.0	94.0	46.7 (21.1)
[Cn] Adv.-New-Detector	14.0	16.2	40.4	39.2	34.7	29.4	17.8	91.0	35.3 (23.2)
[Us] Roberta_dataaug.	4.6	3.6	40.5	7.3	83.1	3.1	5.1	98.8	30.8 (36.8)
[Cn] Adv._Data_Detector	10.1	17.5	27.9	24.8	27.7	28.7	13.5	88.0	29.8 (23.0)
[Lx] Radar R-B CGPT-R	20.0	16.0	4.8	2.5	51.1	62.1	4.4	32.9	24.2 (21.1)
[Ra] Adv. CDMGTD	4.2	3.4	2.1	2.1	6.8	2.9	1.7	2.4	3.2 (1.6)
Average Performance	70.7	66.6	68.4	71.8	74.6	67.7	58.1	78.5	69.5 (5.7)

Subtask B: Performance Across Adversarial Attacks (Official Results)												
	AS	**AD**	**HG**	**IP**	**NS**	**PP**	**MS**	**SY**	**UL**	**WS**	**ZW**	**Total** (σ)
[Le] Leidos v1.0.2	99.2	99.0	**97.3**	98.7	99.2	92.3	98.8	98.6	98.9	99.0	92.7	**97.7 (2.5)**
[Pa] Pangram	99.2	98.7	91.9	**99.3**	99.2	91.6	99.0	96.2	**99.3**	**99.3**	99.3	**97.7** (2.9)
[Le] Leidos v1.0.4	99.1	99.0	94.7	98.7	99.2	94.8	98.8	98.6	98.9	98.8	90.9	97.6 (2.6)
[Le] Leidos v1.0.3	**99.3**	**99.3**	93.6	98.7	**99.4**	**96.3**	**99.2**	**99.1**	99.2	99.2	84.2	97.2 (4.4)
[Le] Leidos v1.0.1	99.0	99.0	86.1	98.1	99.1	94.8	98.9	98.8	98.8	98.5	78.8	95.7 (6.4)
[Us] R-L Focal Loss	97.9	98.2	84.5	93.6	98.1	84.0	97.8	97.4	97.9	97.9	67.1	92.7 (9.5)
[Al] ALERT v1.1	91.8	92.1	68.5	89.7	91.8	57.7	91.0	87.3	91.3	91.2	46.8	82.6 (15.5)
[Lx] Fine-tuned R-B	80.5	78.1	90.4	79.8	79.8	77.9	77.9	74.4	75.0	66.2	**100.0**	80.1 (8.4)
[Al] ALERT v1.2	89.9	89.0	61.9	84.1	88.6	57.1	88.6	84.1	87.2	85.6	40.2	78.8 (16.0)
[Lx] R-B & R-Oai	81.7	79.4	41.7	81.2	81.1	78.1	79.3	75.8	76.1	68.0	86.9	76.0 (11.6)
[Lx] R-Oai & BERT	81.6	79.4	20.9	81.7	82.2	75.8	79.6	77.6	76.7	77.1	83.7	74.9 (17.0)
[Ba] Binoculars	78.2	74.3	37.7	71.7	77.1	80.3	78.0	43.5	73.8	70.1	99.1	71.3 (16.2)
[Ba] Radar	70.8	67.9	59.3	73.7	71.0	67.3	69.5	67.5	70.4	66.1	82.2	69.6 (5.3)
[Mo] MOSAIC-5	72.2	69.5	90.2	73.3	69.7	70.3	71.7	22.7	66.5	67.0	85.5	69.4 (16.3)
[Mo] MOSAIC-4	72.9	70.8	86.6	74.5	71.3	71.9	72.5	28.5	68.6	67.5	71.4	69.3 (13.6)
[Lx] Radar & R-L	70.3	61.2	21.2	73.0	69.9	73.0	63.9	74.9	55.7	60.2	91.3	65.5 (16.6)
[Ba] GLTR	61.2	52.1	24.3	61.4	59.9	47.2	59.8	31.2	48.1	45.8	97.2	53.5 (18.1)
[80] L3-60 Zero-shot	56.6	50.5	3.0	57.4	56.3	50.6	55.6	53.5	57.1	61.9	57.1	51.4 (15.4)
[Ba] openai-roberta-L	52.4	33.2	21.3	55.1	51.7	72.9	39.5	79.4	19.3	40.1	99.9	51.3 (23.6)
[80] M3-60 Zero-shot	55.6	48.6	3.6	56.7	52.2	37.7	53.7	40.2	56.5	59.7	56.5	48.1 (15.4)
[Cn] Adv.-sub.-3	46.7	45.1	20.8	46.7	46.5	18.0	46.8	41.6	46.7	46.7	46.7	41.6 (10.4)
[Cn] Adv.-New-Det.	35.3	35.2	18.9	35.3	35.4	11.9	35.4	31.6	35.3	35.3	35.3	31.7 (7.7)
[Us] Roberta_dataaug.	30.8	31.6	16.4	31.8	30.8	26.8	30.4	30.1	30.8	29.5	11.6	27.6 (6.5)
[Cn] Adv._Data_Det.	29.7	29.4	18.5	29.8	29.6	8.5	29.8	26.9	29.8	29.8	29.8	26.8 (6.5)
[Lx] Radar R-B C-R	22.3	15.2	0.4	4.9	22.0	34.9	18.1	30.0	6.6	4.3	11.0	16.2 (10.6)
[Ra] Adv. CDMGTD	3.2	3.0	24.8	3.2	3.2	3.5	3.2	3.2	3.2	3.2	20.8	6.5 (7.6)
Average Performance	68.4	65.3	49.2	67.4	67.9	60.6	66.8	61.3	64.1	64.2	67.9	64.3 (5.3)

Table 5: TPR at FPR=5% for detectors across different adversarial attacks along with their standard deviation (σ). Baselines are given the [Ba] tag. Abbreviations are: AS: Alternative Spelling, AD: Article Deletion, HG: Homoglyph, IP: Insert Paragraphs, NS: Number Swap, PP: Paraphrase, MS: Misspelling, SY: Synonym Swap, UL: Upper Lower Swap, WS: Whitespace Addition, ZW: Zero-Width Space Addition. Team rankings determined by the highest performing submission (see Table 8).