# It's AI benchmarks

It's AI official website

September 2024

## 1 Introduction

It's AI is an ai-detection tool, which helps people to identify whether text is human-written or ai-generated. In this work we measured an ai-detector from It's AI on several academia-known benchmarks for machine-generated text detectors evaluating and compared our results with other ai-detection tools.

While It's AI returns segmentation predictions for each word, predictions for full text is calculated as an average for words (equal to simple scan functionality on the website).

## 2 Benchmarks

### 2.1 RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. (Paper, Github)

#### 2.1.1 Benchmark description

RAID benchmark is currently the most robust benchmark for ai-detectors evaluation. It addresses the limitations of existing datasets by providing a robust and challenging collection of over 6 million text samples generated by 11 different models across 8 domains, incorporating 11 adversarial attacks and 4 decoding strategies.

The benchmark aims to evaluate the out-of-domain and adversarial robustness of both open-source and closed-source detectors, revealing vulnerabilities in current models and encouraging further research in the field of AI-generated text detection.

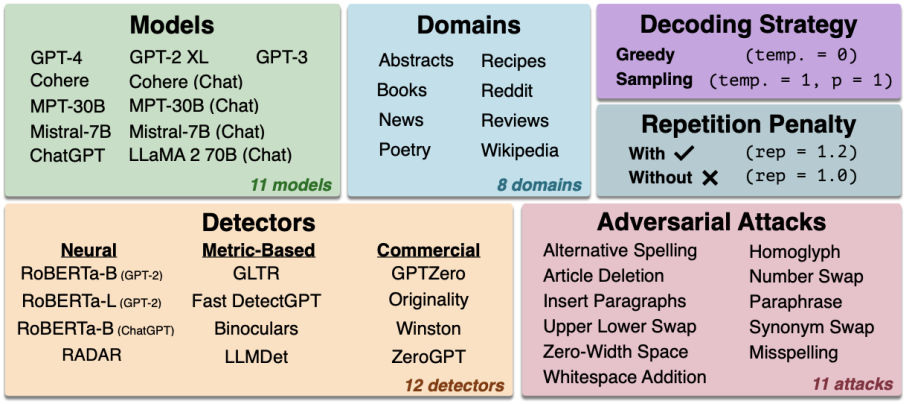


Figure 1: An overview of the structure of the RAID dataset.

The authors generated 2,000 continuations for every combination of domain, model, decoding, penalty, and adversarial attack. This results in roughly 6.2 million generations for testing. Then they evaluated each detector on all pieces of generated text in the dataset.

| Name | Size | Domain coverage? | Model coverage? | Sampling coverage? | Multilingual coverage? | Adversarial coverage? |
|---|---|---|---|---|---|---|
| TuringBench (Uchendu et al., 2021) | 200k | ✗ | ✓ | ✗ | ✗ | ✗ |
| RuATD (Shamardina et al., 2022) | 215k | ✓ | ✓ | ✗ | ✗ | ✗ |
| HC3 (Guo et al., 2023) | 26.9k | ✓ | ✗ | ✗ | ✓ | ✗ |
| MGTBench (He et al., 2023) | 2817 | ✓ | ✓ | ✗ | ✗ | ✓ |
| MULTITuDE (Macko et al., 2023) | 74.1k | ✗ | ✓ | ✗ | ✓ | ✗ |
| AuText2023 (Sarvazyan et al., 2023b) | 160k | ✓ | ✗ | ✗ | ✓ | ✗ |
| M4 (Wang et al., 2023b) | 122k | ✓ | ✓ | ✗ | ✓ | ✗ |
| CCD (Wang et al., 2023a) | 467k | ✗ | ✗ | ✗ | ✓ | ✓ |
| IMDGSP (Mosca et al., 2023) | 29k | ✗ | ✓ | ✗ | ✗ | ✗ |
| HC-Var (Xu et al., 2023) | 145k | ✓ | ✗ | ✗ | ✗ | ✗ |
| HC3 Plus (Su et al., 2024) | 210k | ✓ | ✗ | ✗ | ✓ | ✗ |
| MAGE (Li et al., 2024) | 447k | ✓ | ✓ | ✗ | ✗ | ✗ |
| **RAID (Ours)** | 10M | ✓ | ✓ | ✓ | ✓ | ✓ |

Figure 2: A comparison of the publicly available sources of generated text.

RAID dataset is currenlty the only one that contains a diverse selection of domains, sampling strategies, and adversarial attacks across recent generative models. See RAID ACL 2024 paper for a more detailed comparison.

### 2.1.2 Compared solutions

In the RAID work authors evaluated detectors from three categories: neural, metric-based, and commercial. Neural detectors typically involve fine-tuning a pre-trained language model such as RoBERTa while metric-based detectors typically compute some metric using the output probabilities of an existing generative model. In contrast, commercial detectors tend to provide some doc-

umentation of their performance, but disallow direct access to the models. They tested the following:

1. **Neural:** RoBERTa-Base (GPT2), RoBERTa-Large (GPT2), RoBERTa-Base (ChatGPT), RADAR, E5-small
2. **Metric-Based:** GLTR, Binoculars, Fast DetectGPT, LLMDet
3. **Commercial:** GPTZero, Originality, Winston, ZeroGPT, Desklib

We tested our It's AI solution on closed part of RAID benchmark and compared metrics with the results that RAID authors got on other detectors.

### 2.1.3 Results

RAID benchmark has a leaderboard with a closed test set (without publicly available labels). We have made predictions for it with It's AI detector and submitted them to the leaderboard (instructions for submission can be found in their github).


Figure 3: RAID leaderboard: https://raid-bench.xyz/leaderboard

As for now leaderboard contain both open-sourced solutions and some commercial detectors. It's AI took the first place among them in both cases: scoring samples with adversarial attacks and without them.

To compare our solution with commercial-based detectors, that weren't presented on the leaderboard we took their metrics from the paper. Below we provide table from RAID work extended with It's AI results.

LLMs in the result table are splitted in four groups:

1. **Open-Source:** chat models (llama-c, mistral-c, mpt-c), non-chat models (mistral, mpt, gpt2)

2. **Closed-Source:** chat models (c-gpt, gpt4, cohere), non-chat models (cohere, gpt3)

You can find scores for all considered in the work ai-detectors in Table 1.

| | Open-Source | | | | Closed-Source | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | Chat Models | | Non-Chat Models | | Chat Models | | Non-Chat Models | | |
| | greedy | sampling | greedy | sampling | greedy | sampling | greedy | sampling | |
| **Rep. Penalty?** | ✗ / ✓ | ✗ / ✓ | ✗ / ✓ | ✗ / ✓ | ✗ | ✗ | ✗ | ✗ | - |
| **R-B GPT2** | 84.1 / 52.3 | 77.9 / 26.2 | 98.6 / 44.1 | 60.5 / 35.4 | 70.9 | 41.7 | 65.1 | 52.5 | 59.1 |
| **R-L GPT2** | 79.7 / 41.1 | 71.4 / 19.5 | 98.5 / 43.0 | 67.2 / 53.4 | 61.4 | 34.7 | 61.1 | 48.6 | 56.6 |
| **R-B CGPT** | 80.2 / 63.3 | 75.0 / 39.3 | 53.3 / 26.4 | 14.9 / 1.7 | 59.1 | 38.1 | 46.5 | 39.0 | 44.7 |
| **RADAR** | 88.8 / 77.4 | 85.6 / 66.4 | 91.8 / 63.8 | 48.3 / 31.8 | 81.6 | 75.3 | 72.2 | 67.7 | 70.9 |
| **E5-small (#)** | 99.5 / 98.7 | 99.1 / 96.5 | 98.3 / 96.0 | 85.2 / 89.6 | 96.3 | **94.3** | 85.1 | 79.7 | 93.2 |
| **GLTR** | 89.8 / 67.5 | 83.9 / 38.3 | 99.6 / 56.9 | 75.6 / 63.7 | 80.7 | 54.3 | 75.6 | 63.7 | 70.8 |
| **F-DetectGPT** | 98.6 / 74.5 | 96.2 / 40.5 | 97.8 / 56.1 | 79.7 / 0.6 | 96.0 | 74.1 | 93.8 | 86.3 | 74.5 |
| **LLMDet** | 55.5 / 30.2 | 47.5 / 16.5 | 74.8 / 27.0 | 38.4 / 3.7 | 35.8 | 18.5 | 40.0 | 32.9 | 35.1 |
| **Binoculars** | **99.9** / 86.6 | **99.7** / 60.6 | **99.9** / 62.3 | 72.4 / 0.6 | **99.2** | 92.1 | **99.0** | **95.0** | 80.6 |
| **GPTZero** | 98.8 / 93.7 | 98.4 / 82.5 | 74.7 / 34.6 | 60.6 / 9.4 | 92.3 | 88.5 | 60.6 | 53.4 | 70.6 |
| **Originality** | 98.6 / 86.3 | 97.7 / 72.5 | **99.9** / 64.1 | **89.0** / 51.2 | 96.8 | 89.0 | 91.7 | 85.4 | 85.2 |
| **Winston** | 97.2 / 90.1 | 96.6 / 78.3 | 68.2 / 49.0 | 73.2 / 29.5 | 96.1 | **93.7** | 73.2 | 68.1 | 76.1 |
| **ZeroGPT(*)** | 95.4 / 80.7 | 90.5 / 54.9 | 85.1 / 57.2 | 83.4 / 16.0 | 92.1 | 65.8 | 83.4 | 72.7 | 73.1 |
| **Desklib** | 99.8 / 99.8 | 99.3 / 99.2 | 99.3 / 99.0 | 59.1 / 97.6 | 95.9 | 89.7 | 85.2 | 76.5 | 91.7 |
| **It's AI (sept 2024)** | 99.7 / 99.7 | 99.2 / 99.0 | 99.6 / **99.5** | 54.1 / 96.4 | 95.1 | 87.5 | 87.5 | 79.8 | 91.4 |
| **It's AI** | **99.9** / **99.9** | 99.6 / **99.7** | 99.8 / **99.5** | 76.4 / **98.7** | 96.5 | 93.5 | 87.8 | 79.7 | **94.2** |

Table 1: Accuracy Score at FPR=5% for all detectors across model groups and sampling strategies (no adversarial attacks). Asterisks (*) indicate that the detector was unable to achieve the target FPR. Hashtags (#) indicate that the detector was trained on RAID benchmark train dataset.

It's AI is the best model in 5 out of 12 categories and overall became a new SOTA on the RAID dataset with 1% gap from the second place (E5-small).

| | None | Paraphrase | Synonym | Misspelling | Homoglyph | Whitespace | Delete Articles |
|---|---|---|---|---|---|---|---|
| **R-L GPT2** | 56.7 | 72.9 (+16.2) | 79.4 (+22.7) | 39.5 (-17.2) | 21.3 (-35.4) | 40.1 (-16.6) | 33.2 (-23.5) |
| **RADAR** | 70.9 | 67.3 (-3.6) | 67.5 (-3.4) | 69.5 (-1.4) | 59.3 (-11.6) | 66.1 (-4.8) | 67.9 (-3.0) |
| **GLTR** | 62.6 | 47.2 (-15.4) | 31.2 (-31.4) | 59.8 (-2.8) | 24.3 (-38.3) | 45.8 (-16.8) | 52.1 (-10.5) |
| **Binoculars** | 79.6 | 80.3 (+0.7) | 43.5 (-36.1) | 78.0 (-1.6) | 37.7 (-41.9) | 70.1 (-9.5) | 74.3 (-5.3) |
| **GPTZero** | 66.5 | 64.0 (-2.5) | 61.0 (-5.5) | 65.1 (-1.4) | 66.2 (-0.3) | 66.2 (-0.3) | 61.0 (-5.5) |
| **Originality** | 85.0 | 96.7 (+11.7) | 96.5 (+11.5) | 78.6 (-6.4) | 9.3 (-75.7) | 84.9 (-0.1) | 71.4 (-13.6) |
| **It's AI (sept 2024)** | 91.9 | 75.5 (-16.4) | 79.4 (-12.5) | 91.1 (-0.8) | 74.1 (-17.8) | 92.1 (+0.2) | 85.7 (-6.2) |
| **It's AI** | 94.9 | 83.8 (-11.1) | 99.2 (+4.3) | 93.6 (-1.3) | 65.7 (-29.2) | 33.1 (-61.8) | 90.8 (-4.1) |

Table 2: Accuracy Score at FPR=5% for select detectors across different adversarial attacks. Colors indicate an increase, slight increase, slight decrease, and decrease in performance.

Table 2 shows how different detectors are affected by different adversarial attacks. It's AI scores were taken from leaderboard as an average for all domain, decoding strategies, penalties with different attacks, and that is why they variate from Table 1 (there average was taken by different columns).

All detectors are negatively affected with attacks (on average) and It's AI is not an exception here, but in some cases (like Synonym attack) there is even an improvement when we add this type of attack. We see that there is a field for improvement in homoglyph and whitespace attack and we're going to work on it in next iterations.

## 2.2 CUDRT: Benchmarking the Detection of Human vs. Large Language Models Generated Texts. (Paper, Github)

### 2.2.1 Benchmark description

The CUDRT (Create, Update, Delete, Rewrite, and Translate) dataset is a comprehensive bilingual benchmark designed to evaluate AI-generated text detectors across multiple text generation scenarios in both Chinese and English (we used only English part of it). It includes diverse data sources such as news, theses, community posts, wiki entries, medical data, and financial information, with a total of 81,713 Chinese samples and 197,163 English samples (excluding translation data).

**Tasks in the CUDRT Dataset**

1. **Complete (Create):** This subtask involves generating a complete response or text based on an incomplete prompt.
2. **Polish (Update):** In this subtask, the model is tasked with refining or enhancing existing text. This includes improving grammar, style, and clarity.
3. **Expand (Update):** This operation requires the model to elaborate on a given text, adding more detail or depth.
4. **Summary (Delete):** The summary subtask involves condensing a longer piece of text into a shorter version while preserving the main ideas.
5. **Refine (Delete):** Similar to polishing, this subtask focuses on improving the text's quality by making it more concise or clearer without altering its meaning.
6. **Rewrite (Rewrite):** As mentioned earlier, this subtask requires the model to paraphrase existing text.
7. **Translate (Translate):** This subtask involves converting text from one language to another.
8. **Question-Answering (Create):** In this subtask, the model is provided with a text passage and asked to answer specific questions based on the content.

### 2.2.2 Compared solutions

In the cross-dataset detection section of the paper "CUDRT: Benchmarking the Detection of Human vs. Large Language Models Generated Texts," three AI-generated text detectors were evaluated. These detectors were selected based on their common usage and recent advances in the field. They were previously

trained using HC3 dataset and tested on a new dataset created for this study without any fine-tuning.

#### Multiscale Positive-Unlabeled

The MPU reformulates AI-generated text detection as a partial Positive-Unlabeled (PU) problem, using a length-sensitive Multiscale PU loss function and an abstract recurrent model to estimate prior probabilities across different text lengths. It includes a Text Multiscaling module that generates texts of varying lengths through random sentence deletion, enhancing detection of short texts while maintaining effectiveness for longer texts.

#### RoBERTa

The RoBERTa classifier is a fine-tuned text classification tool based on the pre-trained RoBERTa model, specifically designed to detect AI-generated text. It improves performance on various natural language processing tasks by enhancing data handling and model configuration. The classifier distinguishes between human and AI-generated texts, focusing on both Chinese and English languages.

#### XLNet

XLNet utilizes a permutation language model to consider all possible combinations of word order during training, addressing inconsistencies between pre-training and fine-tuning in BERT. It incorporates Transformer-XL technology for better handling of long-range dependencies and learns to identify AI involvement in text generation by extracting features from both human- and AI-generated texts.

### 2.2.3 Results

For scoring It's AI we used DatasetFinal from CUDRT official github.

| Model | It's AI | It's AI (sept 2024) | MPU | RoBERTa | XLNet |
|---|---|---|---|---|---|
| Baichuan | 0.500 | 0.551 | **0.601** | 0.524 | 0.409 |
| ChatGLM | **0.912** | 0.892 | 0.670 | 0.520 | 0.401 |
| GPT3.5 | **0.780** | 0.762 | 0.658 | 0.526 | 0.409 |
| Llama2 | 0.708 | 0.716 | **0.738** | 0.537 | 0.414 |
| Llama3 | **0.879** | 0.854 | 0.740 | 0.528 | 0.408 |

Table 3: F1-scores for different models

As shown in Table 3 It's AI significantly outperform all three models tested in CUDRT work on out of domain validation and achieve average f1-score 0.755. The hardest to detect model was Baichuan, which is trained on a combined

corpose of Chinese and English texts and have Chinese hallucinations in eglish texts, which english-only It's AI solution couldn't handle properly.

| Task | It's AI | It's AI (sept 2024) | MPU | RoBERTa | XLNet |
|---|---|---|---|---|---|
| Complete | 0.814 | **0.820** | 0.744 | 0.451 | 0.342 |
| Expand | **0.892** | 0.850 | 0.698 | 0.388 | 0.340 |
| Polish | **0.864** | 0.800 | 0.627 | 0.387 | 0.340 |
| QA | 0.941 | 0.901 | **0.947** | 0.660 | 0.641 |
| Refine | **0.886** | 0.848 | 0.637 | 0.385 | 0.340 |
| Rewrite | **0.911** | 0.862 | 0.664 | 0.383 | 0.337 |
| Summary | 0.741 | 0.759 | 0.657 | **0.924** | 0.520 |
| Translate | 0.469 | 0.505 | 0.637 | **0.643** | 0.405 |

Table 4: F1-scores for different tasks excluding Baichuan model

We think that the reason for a low detection quality on CUDRT dataset among all models is because of the structure of dataset: in many tasks ai-generated texts are not fully ai-generated - they have a human basis text and then change it instead of writting from scratch (for example tasks Expand, Polish, Refine, Summary and Translate).

The easiest for detection was a classic QA task, when with a given prompt LLM write a text, while the hardest one was a translation, when LLM doesn't have a space for creation and just translate a human-written text.

Overall, It's AI outperformed other approaches in 5/8 tasks and took the second place in 2/8 tasks.

## 2.3 GPT-generated Text Detection: Benchmark Dataset and Tensor-based Detection Method. (Paper, Github)

### 2.3.1 Benchmark description

The dataset presented in the paper "GPT-generated Text Detection: Benchmark Dataset and Tensor-based Detection Method" is called the GPT Reddit Dataset (GRiD). This dataset is specifically designed to evaluate the performance of detection models in identifying text generated by ChatGPT.

GRiD consists of context-prompt pairs sourced from Reddit, featuring responses generated by humans and generated by ChatGPT (GPT-3.5 was used). Total dataset size is 6500 samples.

### 2.3.2 Compared solutions

Authors tested three distinct models — Random Forest, Support Vector Machine (SVM) and BERT to assess their efficacy in GPT-generated text detec-

tion. For both SVM and Random Forest a simple TF-IDF vectorizer was used to transform text data into numerical data. Selected models **were trained** on the dataset and their metrics were obtained via 10-fold cross-validation.

#### 2.3.3 Results

For scoring It's AI we used reddit filtered dataset from GriD official github.

| | **F1** | **AUC** |
|---|---|---|
| **BERT** | 0.934 | 0.984 |
| **SVM** | 0.813 | 0.845 |
| **Random Forest** | 0.787 | 0.825 |
| **It's AI (sept 2024)** | 0.955 | 0.992 |
| **It's AI** | **0.973** | **0.998** |

Table 5: F1 and AUC scores for GriD benchmark

Despite the fact that for It's AI this validation was out-of-domain, while other covered in the work detectors were fine-tuned on the dataset, It's AI outperformed all of them and took top-1 solution with f1-score 0.973 and AUC-score 0.998.

## 3 Conclusion

In this work we measured ai-detector from It's AI on three benchmarks:

1. **RAID.** The most diversed and representable benchmark up to September 2024 with more than 10m samples.
2. **CUDRT.** Specifies on modifying human-written text with several tasks.
3. **GriD.** Consist of reddit context-prompt pairs and completion for them.

On RAID dataset we got accuracy 94.2% on non-attacked texts at FPR 5% and outperformed all other detectors - second place was e5-small with 93.2%. All results were obtained from the official leaderboard where we submitted predictions on the test set. So, It's AI is officially a new SOTA on RAID benchmark!

In the CUDRT dataset only open-source solutions were considered, and we took first place in 5 out of 8 tasks and on average outperformed the best considered solution on 7% of f1-score.

GriD paper contained comparison of a few trained on the data models (while for us it was out of domain validation), but It's AI was able to outperform them anyway (97% f1-score vs 93% of best solution from the work) and become new SOTA on GriD dataset as well.

Overall, It's AI become a new SOTA on all three considered benchmarks and showed an impressive quality on out of domain validation. These are very promising results and we are hoping to make them even better in the future.