Quick overview:
The “Binoculars” Paper

Fede Nanni

Automatic detection of LLM Generated Text

Automatic detection of LLM Generated Text

1. Watermarking

Automatic detection of LLM Generated Text

1. Watermarking
2. Fine-tuning a language model for the binary detection task

Automatic detection of LLM Generated Text

1. Watermarking
2. Fine-tuning a language model for the binary detection task
3. Zero-shot based on statistical signature

Automatic detection of LLM Generated Text

Posted by u/MT1699 14 days ago

AI tools to detect LLM generated text

Question | Help

Automatic detection of LLM Generated Text

Posted by u/MT1699 14 days ago

AI tools to detect LLM generated text

Question | Help

stilllllinit · 14 days ago

It's impossible.

25 Reply Share ...

Automatic detection of LLM Generated Text

You

is it possible to do automatic detection of LLM generated text without any training data or watermarking?

‹ 2 / 2 ›

ChatGPT

Detecting whether text was generated by a language model (LLM) like GPT-3 without training data or watermarking is challenging. However, there are some techniques that can be employed to identify generated text, although they may not be foolproof:

1. **Statistical Analysis**: Analyzing statistical properties of the text, such as word frequency distributions, sentence lengths, or syntactic structures, can sometimes reveal patterns indicative of generated text. However, this approach may not be reliable as it heavily relies on assumptions about the characteristics of the generated text.
2. **Pattern Recognition**: Look for specific patterns or anomalies that are common in generated text. For instance, language models often exhibit repetition or inconsistency in longer passages, and they may struggle with maintaining coherence over multiple paragraphs.
3. **Contextual Understanding**: Compare the text against a large corpus of human-written text to see if it fits within the context and style of natural language. Language models can sometimes produce text that is slightly off or unnatural when scrutinized closely.
4. **Semantic Analysis**: Analyze the meaning and coherence of the text. Generated text may lack deep semantic understanding or logical consistency, leading to inconsistencies or nonsensical

Automatic detection of LLM Generated Text

fedenanni 2 months ago

Hello folks! I was wondering if you could point me to any text classifier you are aware of for detecting if a string of text is llm generated. I would like to set up a small benchmark for these tools - it's ok both if they are demos from a company (like: https://originality.ai/ai-checker) or tools described in a paper. Thank you all in advance!

Note that this one from OpenAI doesn't exist anymore: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

👀 3

4 replies

Sanjay 2 months ago

https://arxiv.org/abs/2401.12070

> **arXiv.org**
>
> **Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text**
>
> Detecting text generated by modern large language models is thought to be hard, as both LLMs and humans can exhibit a wide range of complex behaviors. However, we find that a score based on contrasting two closely related language models is highly accurate at separating human-generated and machine-generated text. Based on this mechanism, we propose a novel LLM detector that only requires simple calculations using a pair of pre-trained LLMs. The method, called Binoculars, achieves state-of-the-art accuracy without any training data. It is capable of spotting machine text from a range of modern LLMs without any model-specific modifications. We comprehensively evaluate Binoculars on a number of... Show more

🙌 1

The paper

~~Submitted~~ to ICLR 2024

Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text

Abhimanyu Hans*
University of Maryland

Avi Schwarzschild*
Carnegie Mellon University

Valeriia Cherepanova
University of Maryland

Hamid Kazemi
University of Maryland

Aniruddha Saha
University of Maryland

Micah Goldblum
New York University

Jonas Geiping
ELLIS Institute & MPI for Intelligent Systems,
Tübingen AI Center

Tom Goldstein
University of Maryland

The code

README BSD-3-Clause license

Binoculars: Zero-Shot Detection of LLM-Generated Text [paper][demo]

The code

```
$ git clone https://github.com/ahans30/Binoculars.git
$ cd Binoculars
$ pip install -e .
```

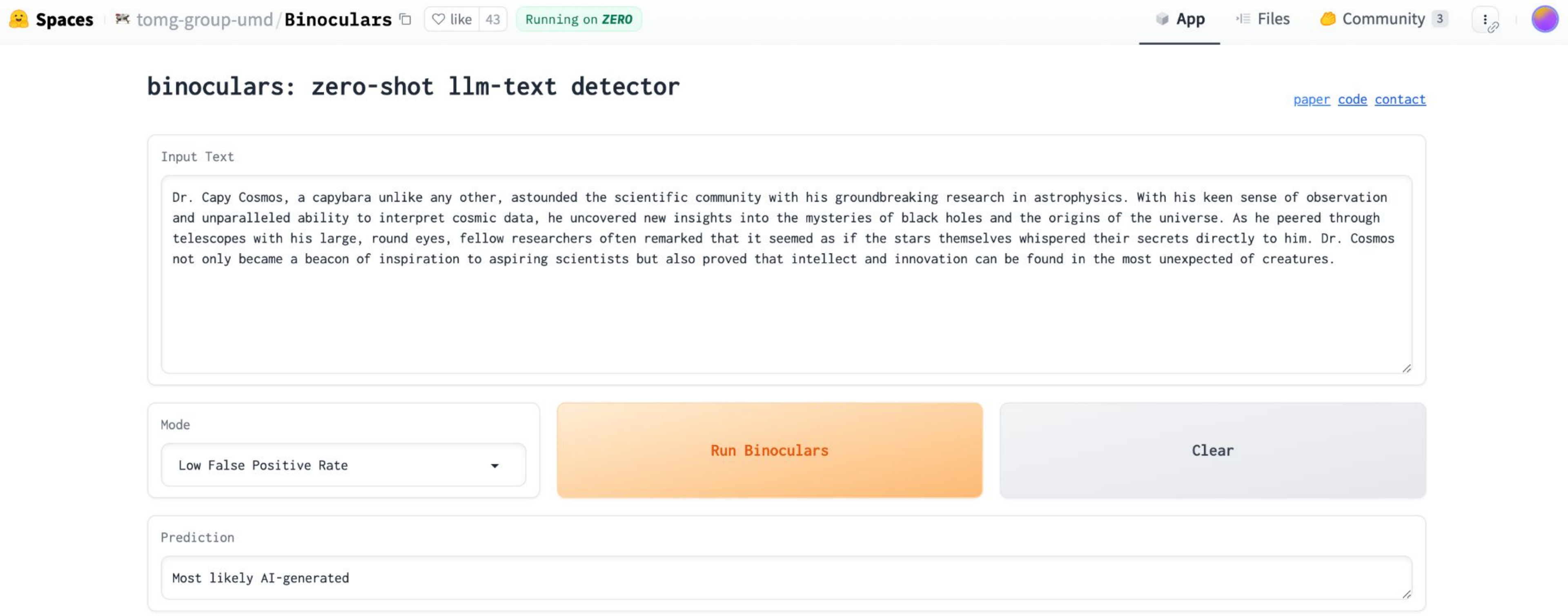
```
from binoculars import Binoculars

bino = Binoculars()

# ChatGPT (GPT-4) output when prompted with “Can you write a few sentences about a capybara th
sample_string = '''Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific comm
groundbreaking research in astrophysics. With his keen sense of observation and unparalleled a
cosmic data, he uncovered new insights into the mysteries of black holes and the origins of th
peered through telescopes with his large, round eyes, fellow researchers often remarked that i
stars themselves whispered their secrets directly to him. Dr. Cosmos not only became a beacon
aspiring scientists but also proved that intellect and innovation can be found in the most une

print(bino.compute_score(sample_string))  # 0.75661373
print(bino.predict(sample_string))  # 'Most likely AI-Generated'
```

The demo


Spaces
tomg-group-umd/Binoculars
like 43
Running on ZERO
App
Files
Community 3
binoculars: zero-shot llm-text detector
paper code contact
Input Text
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe. As he peered through telescopes with his large, round eyes, fellow researchers often remarked that it seemed as if the stars themselves whispered their secrets directly to him. Dr. Cosmos not only became a beacon of inspiration to aspiring scientists but also proved that intellect and innovation can be found in the most unexpected of creatures.
Mode
Low False Positive Rate
Run Binoculars
Clear
Prediction
Most likely AI-generated

How does it work?

How does it work?

$$B_{\mathcal{M}_1}(s) = \log \mathrm{PPL}_{\mathcal{M}_1}(s)$$

Language models (here, M1) produce low-perplexity text relative to humans

The capybara problem

"Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe. As he peered through telescopes with his large, round eyes, fellow researchers often remarked that it seemed as if the stars themselves whispered their secrets directly to him. Dr. Cosmos not only became a beacon of inspiration to aspiring scientists but also proved that intellect and innovation can be found in the most unexpected of creatures." - ChatGPT

Figure 2: This quote is LLM output from ChatGPT (GPT-4) when prompted with "Can you write a few sentences about a capybara that is an astrophysicist?" The Falcon LLM assigns this sample a high perplexity (2.20), well above the mean for both human and machine data. Despite this problem, our detector correctly assigns a *Binoculars* score of 0.73, which is well below the global threshold of 0.901, resulting in a correct classification with high confidence. For reference, DetectGPT wrongly assigns a score of 0.14, which is below it's optimal threshold of 0.17, and classifies the text as human. GPTZero assigns a 49.71% score that this text is generated by AI.

How does it work?

$$B_{\mathcal{M}_1,\mathcal{M}_2}(s) = \frac{\log \mathrm{PPL}_{\mathcal{M}_1}(s)}{\log \mathrm{X\text{-}PPL}_{\mathcal{M}_1,\mathcal{M}_2}(s)}$$

Numerator: Language models produce low-perplexity text relative to humans

Denominator: cross-perplexity, to see whether the tokens that appear in a string are surprising relative to the baseline perplexity of another LLM acting on the same string.

How does it work?

$$B_{\mathcal{M}_1,\mathcal{M}_2}(s) = \frac{\log \mathrm{PPL}_{\mathcal{M}_1}(s)}{\log \mathrm{X\text{-}PPL}_{\mathcal{M}_1,\mathcal{M}_2}(s)}$$

Numerator: Language models produce low-perplexity text relative to humans

Denominator: cross-perplexity, to see whether the tokens that appear in a string are surprising relative to the baseline perplexity of another LLM acting on the same string.

M_1= Falcon-7b - M_2= Falcon-7b-instruct

Evaluation

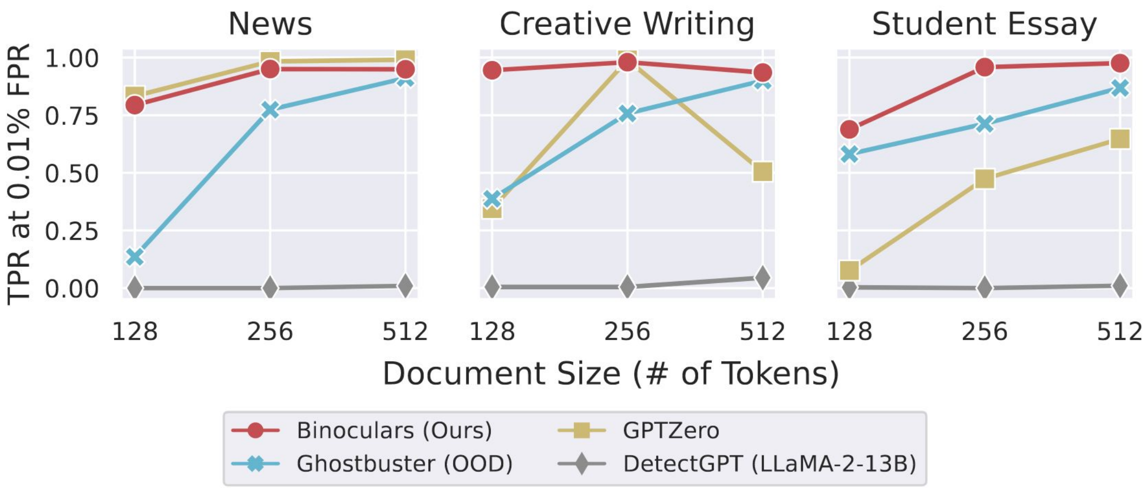


Figure 3: **Impact of Document Size on Detection Performance**. The plot displays the TPR at 0.01% FPR across varying document sizes. The x-axis represents the number of tokens of the observed document, while the y-axis indicates the corresponding detection performance, highlighting the *Binoculars* ability to detect with a low number of tokens.

Evaluation

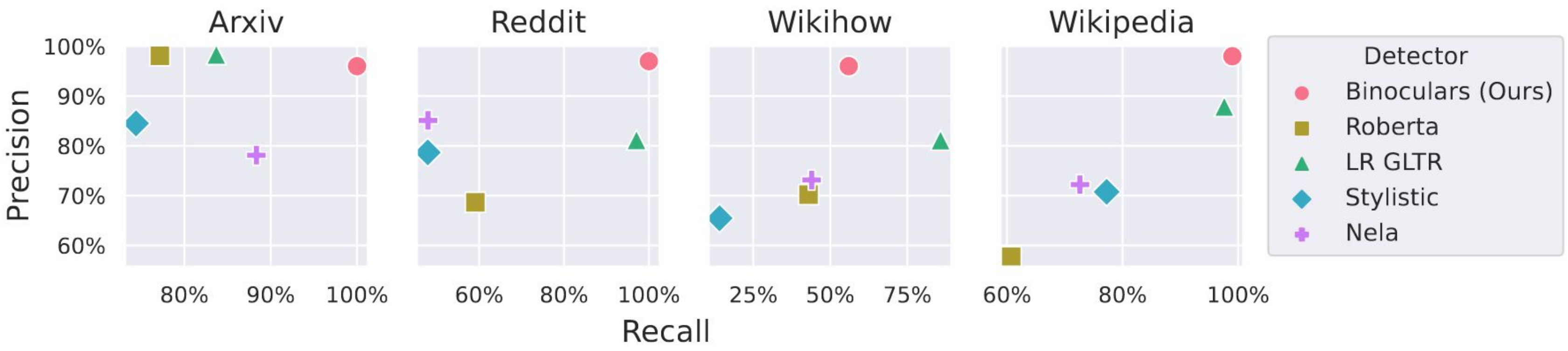


Figure 5: **Detection of ChatGPT-generated text in various domains from M4 Dataset.** Binoculars maintain high precision over 4 domains using the global threshold (tuned out-of-domain) for detection. We use the mean of out-of-domain performance metrics reported by Wang et al. (2023)

Evaluation

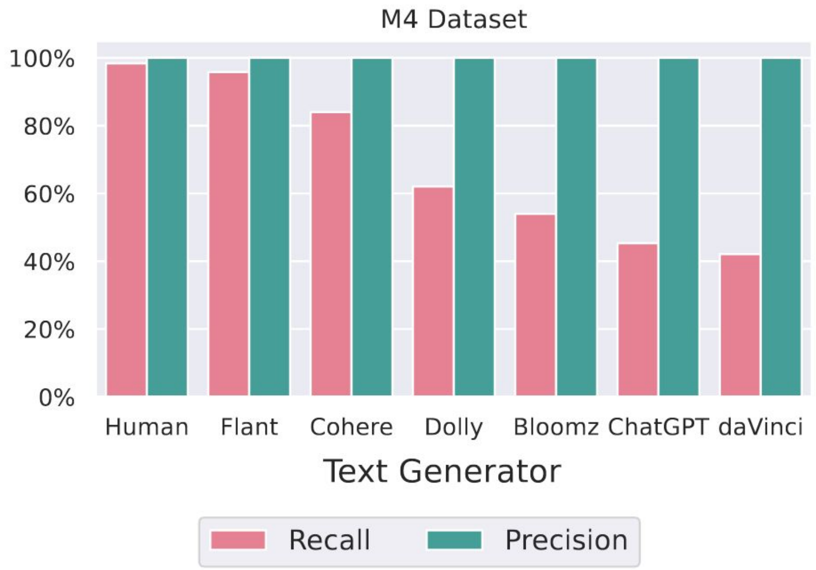


Figure 6: Performance of *Binoculars* on samples from various generative models.

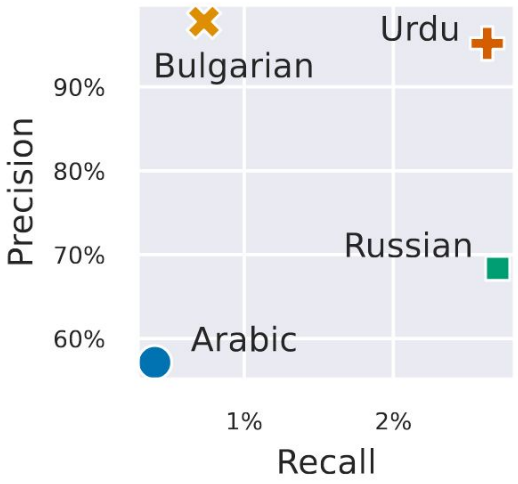


Figure 7: *Binoculars* operates at high precision in Bulgarian and Urdu, but with low recall in all four languages.

Summary of interesting things

1. Very simple idea
2. Relatively easy to modify the used LLMs
3. Quick to test and see whether it works in your setting
    a. It worked well in ours, where we wanted to see whether text created by a LLM could be identified as such by Binoculars
4. They highlight an underlying similarity between modern LLMs, as they all use **nearly identical transformer components** and are likely trained on datasets comprising mostly **Common Crawl data** from similar time periods.

Questions?

Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text

Abhimanyu Hans* University of Maryland

Avi Schwarzschild* Carnegie Mellon University

Valeriia Cherepanova University of Maryland

Hamid Kazemi University of Maryland

Aniruddha Saha University of Maryland

Micah Goldblum New York University

Jonas Geiping ELLIS Institute & MPI for Intelligent Systems, Tübingen AI Center

Tom Goldstein University of Maryland