

# Can AI Generated Text be detected or Not? – A comparative Study of Tools

Group 14

Abhilash Rajesh Bagalkoti  
Vismaya Maliye  
Prajwal Manjunath

# Introduction

- The rapid progress of Large Language Models (LLMs) has made them capable of performing astonishingly well on various tasks.
- The unregulated use of these models, however, can potentially lead to malicious consequences such as plagiarism, generating fake news, spamming, etc.
- Therefore, reliable detection of AI-generated text can be critical to ensure the responsible use of LLMs
- We have worked on comparing the performance of current detector models.

# Detectors of AI generated text

- The categories of existing algorithms for detecting machine generated text are :
  - Watermarking algorithms
    - A “watermark” is a modification to generated text that can be detected by algorithm while remaining unchanged to human readers.
    - They are difficult to remove and have little effect on the quality of generation.
    - “[A Watermark for Large Language Models](#)” : This paper proposes a simple algorithm that only requires access to LLM logits at each time stamp to add watermarks.

## No watermark

Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)  
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet

## With watermark

- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

- Statistical Outlier detection methods

- There is no modification to the generative algorithm like watermarking.
- Earlier methods detect irregularities in measures such as entropy, ngram, perplexity etc.
- Recent tools after release of ChatGPT:
  - [GPTZero](#)
  - [Detect- GPT](#) (Open-Source Tool)
- DetectGPT uses an observation that model-generated text lies in the negative curvature regions of the model's log probability function.
- Perturbations are generated using : T5 ( we use "[t5-small](#)" )
- Text is classified as model generated if log probability of unperturbed text is significantly higher than perturbations.

- Classifiers:

- Classifiers are finetuned to distinguish human written text from machine generated text.
- Recently OpenAI fine-tuned a GPT model to perform this discrimination task and released it as a web interface.
- They fine-tuned this classifier using generations from 34 language models, with text sourced from Wikipedia, WebText, and their internal data.
- <https://platform.openai.com/ai-text-classifier>

# Text Generation Models Used:

- GPT2-medium (<https://huggingface.co/gpt2-medium>)
  - Number of parameters: 335M
  - Model Size: 1.52GB
- Facebook/OPT-1.3b (<https://huggingface.co/facebook/opt-1.3b>)
  - Number of parameters: 1.3B
  - Model Size: 2.63 GB
- EleutherAI/gpt-neo-1.3B (<https://huggingface.co/EleutherAI/gpt-neo-1.3B>)
  - Number of parameters: 1.3B
  - Model Size: 5.31 GB
- GPT-3.5-17.5B davinci-003

# Task - Paraphrasing

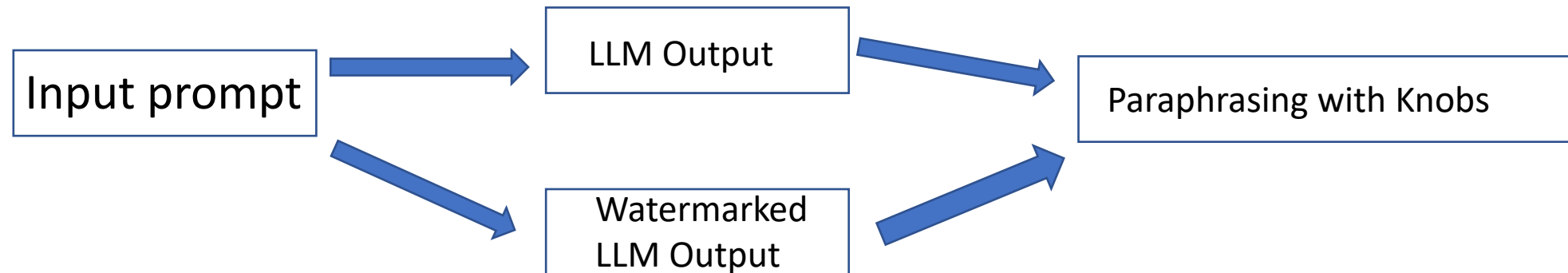
- The 2 main factors which classify the text as good phrase are
  - If the generated text conveys the same meaning as the original
  - If the text is grammatically/fluent correct English.
- We use [prithivida/parrot paraphraser on T5](#) as it contains knobs to control adequacy, fluency and diversity.

	0.0	0.04	0.08	0.16	0.25
Knob value	1.0	0.96	0.92	0.84	0.75

- Span replacement vs Paraphrasing.

# Evaluation metric and Dataset Used

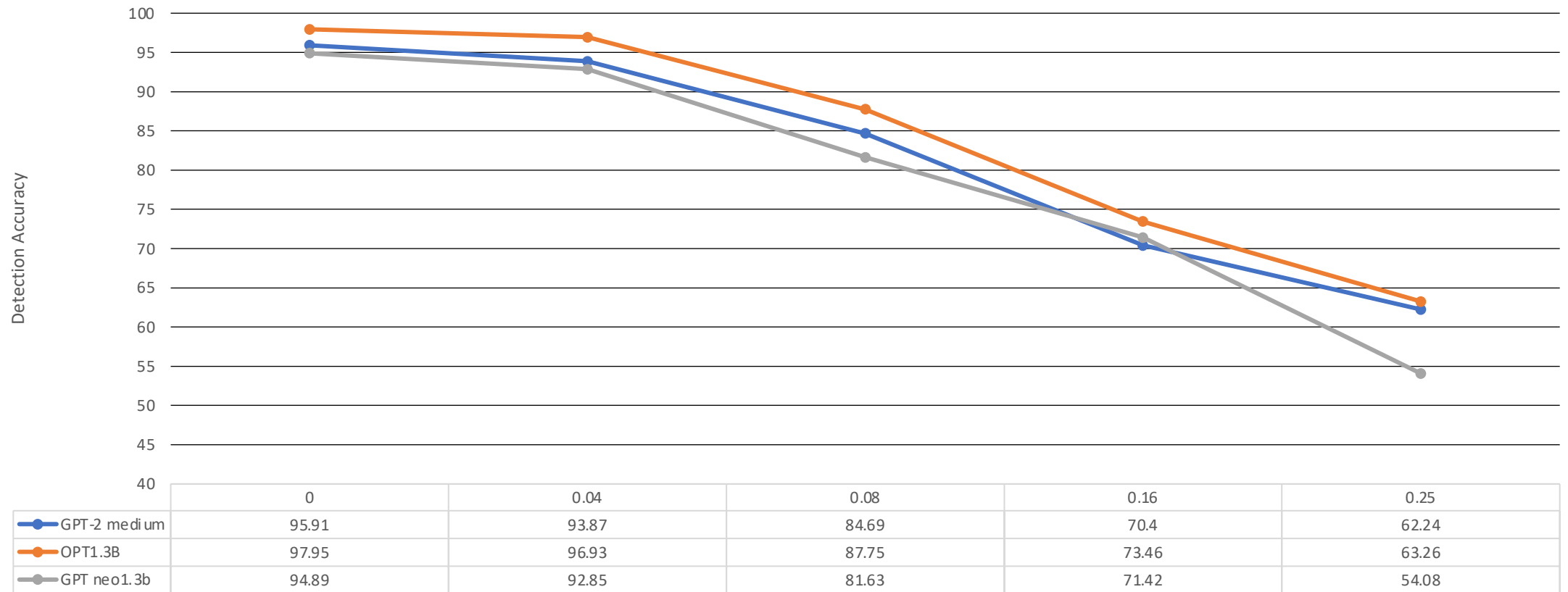
- Evaluation metric : Detection accuracy.
- Dataset : [Xsum](#)
  - We use the first 200 examples in Test set (length of each example is >200 )
  - Length of Input prompt used for each model : 100  
(Detect-gpt and other papers used 50-60)
  - No condition is given on the number of tokens generated from the model.



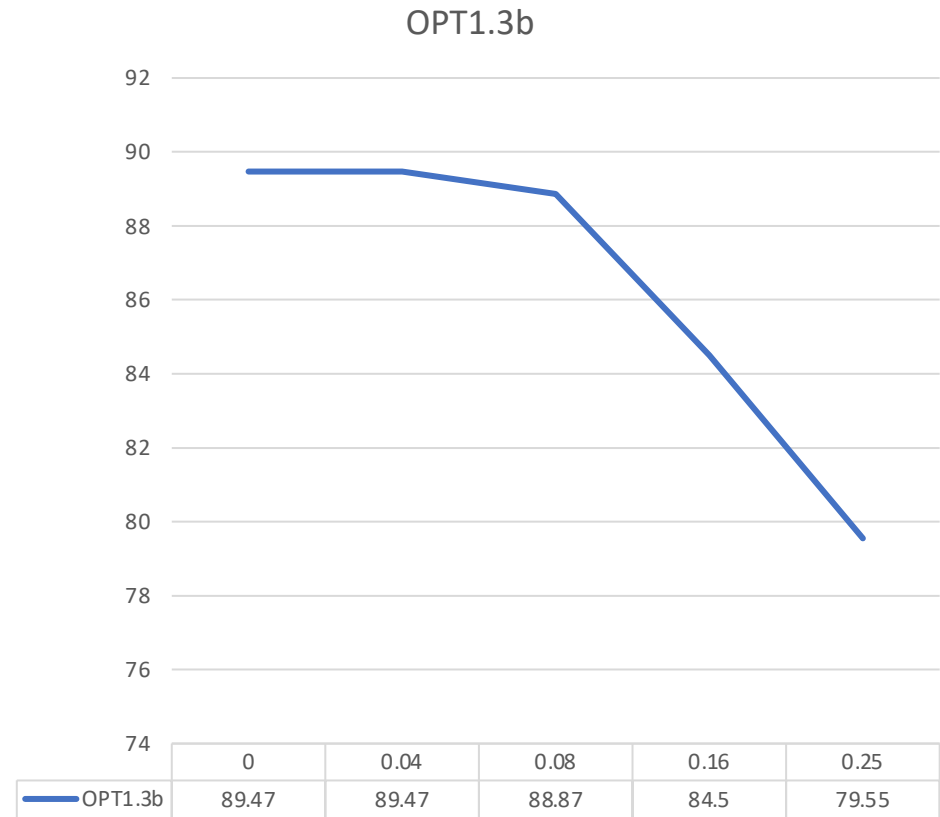
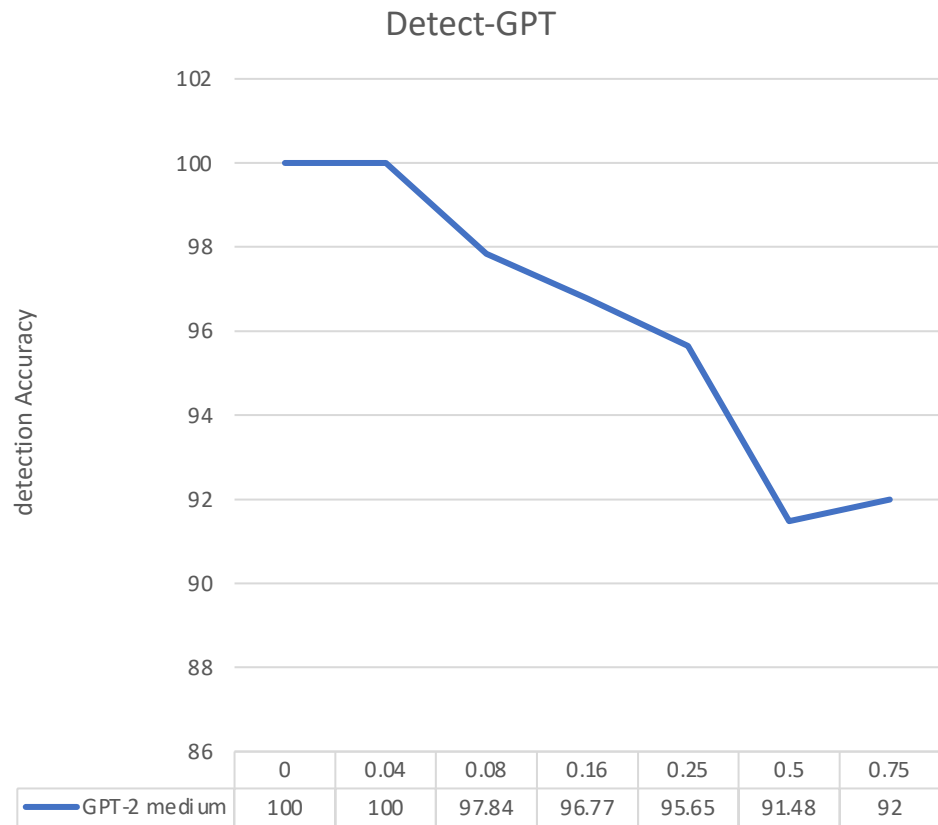


# Results: Watermark Detection

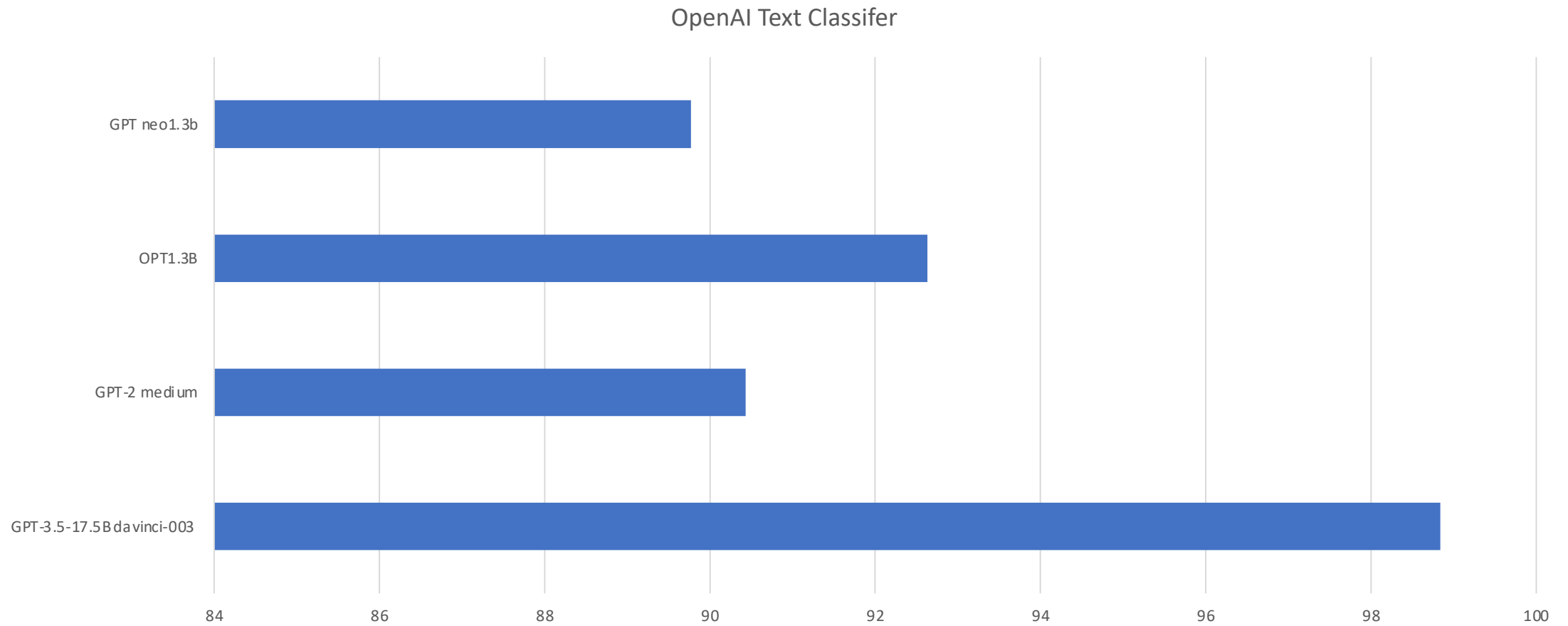
Watermark Detection - Paraphrasing



# Results: Detect-GPT



# Result: OpenAI Text Classifier



# Conclusion

- The experiments show that paraphrasing of LLM outputs helps to evade the detectors.
- Size of LLM models for generation and detection would affect the performance. For a sufficiently large model, best detector can only perform better than random classifier.
- Watermarking-based detectors can be spoofed to make human-composed text detected as watermarked.

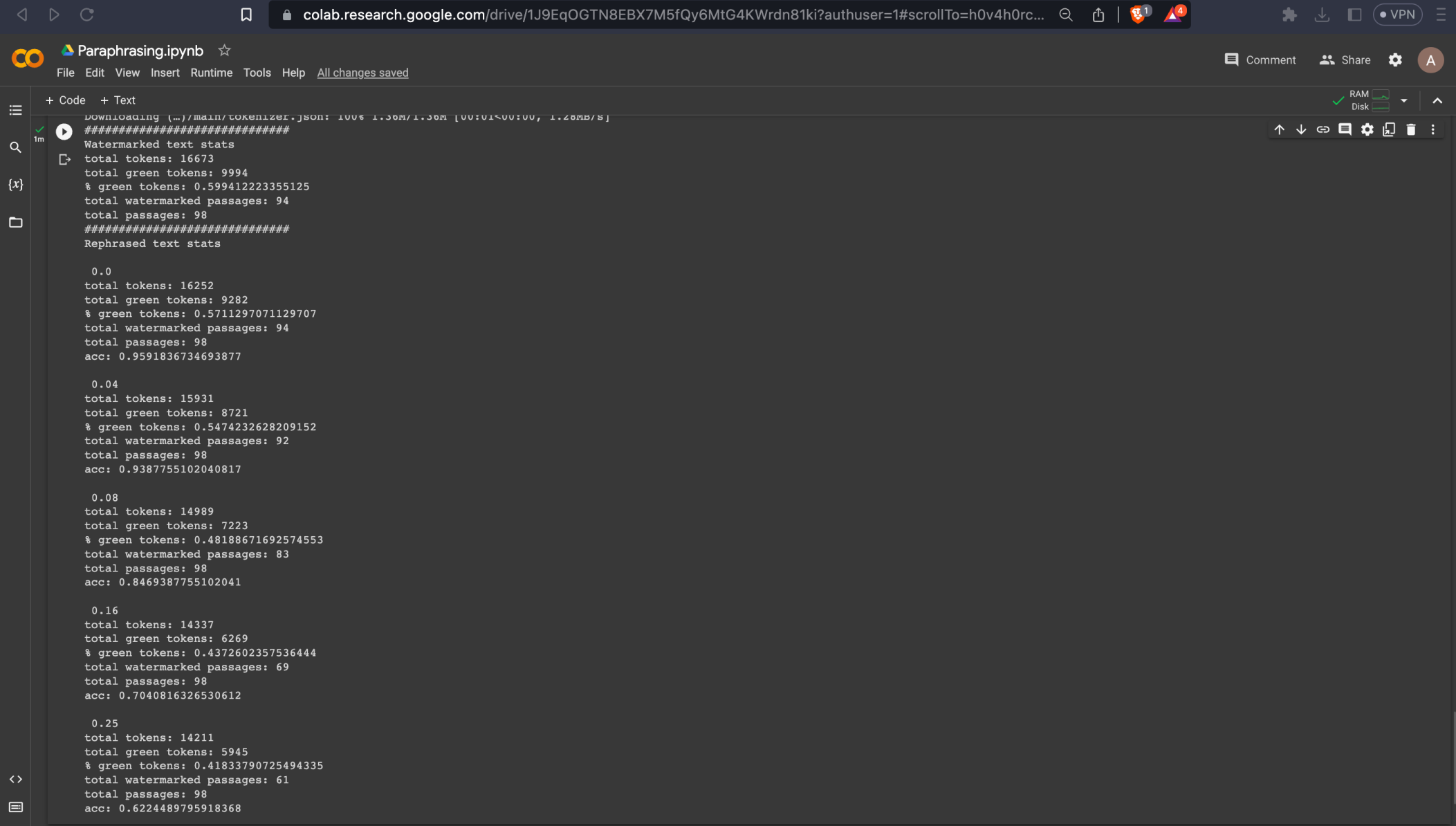
# Future Scope:

- Multiple times paraphrasing effect
- Model used for paraphrasing is T5-small, Use bigger models of T5 or Pegasus
- Study on False Positive Rate
- New methods of detection
- Emoji Attack on GPT



Thank You

# Result - Screenshots



```
Downloading (...) /main/tokenizer.json: 100% 1.36M/1.36M [00:01<00:00, 1.28MB/s]
#####
Watermarked text stats
total tokens: 16673
total green tokens: 9994
% green tokens: 0.599412223355125
total watermarked passages: 94
total passages: 98
#####
Rephrased text stats

0.0
total tokens: 16252
total green tokens: 9282
% green tokens: 0.5711297071129707
total watermarked passages: 94
total passages: 98
acc: 0.9591836734693877

0.04
total tokens: 15931
total green tokens: 8721
% green tokens: 0.5474232628209152
total watermarked passages: 92
total passages: 98
acc: 0.9387755102040817

0.08
total tokens: 14989
total green tokens: 7223
% green tokens: 0.48188671692574553
total watermarked passages: 83
total passages: 98
acc: 0.8469387755102041

0.16
total tokens: 14337
total green tokens: 6269
% green tokens: 0.4372602357536444
total watermarked passages: 69
total passages: 98
acc: 0.7040816326530612

0.25
total tokens: 14211
total green tokens: 5945
% green tokens: 0.41833790725494335
total watermarked passages: 61
total passages: 98
acc: 0.6224489795918368
```

- detectgpt
- sample\_data
- detect\_gpt\_exp.py
- llm\_output.pkl
- rephrased\_org\_0.00.pkl
- rephrased\_org\_0.04.pkl
- rephrased\_org\_0.08.pkl
- rephrased\_org\_0.16.pkl
- rephrased\_org\_0.25.pkl

```
python detect_gpt_exp.py
2023-05-04 21:46:26.955132: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
Loading Base model of type gpt2-medium ...
Loading mask model of type t5-small ...

0.0
 2% 2/100 [00:18<14:06, 8.64s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
12% 12/100 [01:44<11:39, 7.95s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
25% 25/100 [03:46<11:23, 9.11s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
42% 42/100 [05:53<07:30, 7.77s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
46% 46/100 [06:29<07:35, 8.43s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
48% 48/100 [06:43<06:34, 7.59s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
67% 67/100 [09:23<04:24, 8.01s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
72% 72/100 [10:07<03:47, 8.12s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
92% 92/100 [12:36<00:50, 6.30s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [13:49<00:00, 8.30s/it]
knob: 0.0
Total processed data: 94
Detection Accuracy with: 1.0

0.04
 3% 3/100 [00:23<12:44, 7.88s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
25% 25/100 [03:38<11:20, 9.08s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
WARNING: 1 texts have no fills. Trying again [attempt 1].
32% 32/100 [04:31<08:35, 7.58s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
58% 58/100 [07:50<05:22, 7.68s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
77% 77/100 [10:26<02:55, 7.61s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
87% 87/100 [11:45<01:51, 8.54s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
92% 92/100 [12:11<00:51, 6.40s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
96% 96/100 [12:44<00:31, 7.91s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [13:20<00:00, 8.01s/it]
knob: 0.04
Total processed data: 94
Detection Accuracy with: 1.0

0.08
 0% 0/100 [00:00<?, ?it/s]WARNING: 1 texts have no fills. Trying again [attempt 1].
15% 15/100 [02:01<13:18, 9.39s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
17% 17/100 [02:24<14:08, 10.22s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
27% 27/100 [03:41<07:56, 6.52s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
35% 35/100 [04:40<07:48, 7.20s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
42% 42/100 [05:23<06:30, 6.72s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
54% 54/100 [07:01<06:11, 8.08s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
92% 92/100 [11:38<00:46, 5.83s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [12:45<00:00, 7.65s/it]
knob: 0.08
Total processed data: 93
Detection Accuracy with: 0.978494623655914

0.16
```

RAM Disk



Files

- detectgpt
- sample\_data
- detect\_gpt\_exp.py
- llm\_output.pkl
- rephrased\_org\_0.00.pkl
- rephrased\_org\_0.04.pkl
- rephrased\_org\_0.08.pkl
- rephrased\_org\_0.16.pkl
- rephrased\_org\_0.25.pkl

```
+ Code + Text
[11] 27% 27/100 [03:41<07:56, 6.52s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
35% 35/100 [04:40<07:48, 7.20s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
42% 42/100 [05:23<06:30, 6.72s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
54% 54/100 [07:01<06:11, 8.08s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
92% 92/100 [11:38<00:46, 5.83s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [12:45<00:00, 7.65s/it]
knob: 0.08
Total processed data: 93
Detection Accuracy with: 0.978494623655914

0.16
18% 18/100 [02:20<11:06, 8.12s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
30% 30/100 [03:51<08:12, 7.04s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
34% 34/100 [04:18<07:19, 6.66s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
40% 40/100 [04:52<05:11, 5.19s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
WARNING: 1 texts have no fills. Trying again [attempt 1].
58% 58/100 [07:13<04:51, 6.94s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
61% 61/100 [07:38<04:54, 7.55s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
WARNING: 1 texts have no fills. Trying again [attempt 1].
70% 70/100 [08:49<04:04, 8.14s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
79% 79/100 [09:58<02:39, 7.60s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [12:23<00:00, 7.44s/it]
knob: 0.16
Total processed data: 93
Detection Accuracy with: 0.967741935483871

0.25
2% 2/100 [00:13<10:23, 6.36s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
4% 4/100 [00:27<10:23, 6.49s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
6% 6/100 [00:43<11:55, 7.61s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
9% 9/100 [01:08<11:58, 7.90s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
20% 20/100 [02:38<11:32, 8.66s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
23% 23/100 [03:02<10:24, 8.11s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
24% 24/100 [03:11<10:45, 8.50s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
28% 28/100 [03:35<08:09, 6.80s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
38% 38/100 [04:41<06:04, 5.88s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
55% 55/100 [06:44<05:42, 7.61s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
58% 58/100 [07:00<04:32, 6.48s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
WARNING: 1 texts have no fills. Trying again [attempt 1].
63% 63/100 [07:35<03:51, 6.26s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
76% 76/100 [09:19<02:50, 7.09s/it]WARNING: 1 texts have no fills. Trying again [attempt 1].
100% 100/100 [11:59<00:00, 7.19s/it]
knob: 0.25
Total processed data: 92
Detection Accuracy with: 0.9565217391304348
```

RAM Disk