# Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature


**Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, Yue Zhang**

(baoguangsheng@westlake.edu.cn)


Zhejiang University

Westlake University

Shanghai Polytechnic University

Nanyang Technological University

# CONTENTS


**01**  **Background**

**02**  Fast-DetectGPT

**03**  Experiments

**04**  Code and Demo


# Motivation

西 湖 大 學
WESTLAKE UNIVERSITY

**Which news is fake?**

**News 1:** Maj Richard Scott, 40, is accused of driving at speeds of up to 95mph (153km/h) in bad weather before <u>the smash on a B-road in Wiltshire. Gareth Hicks, 24, suffered fatal injuries when the van he was asleep in was hit by Mr Scott's Audi A6. Maj Scott denies a charge of causing death by careless driving ...</u>

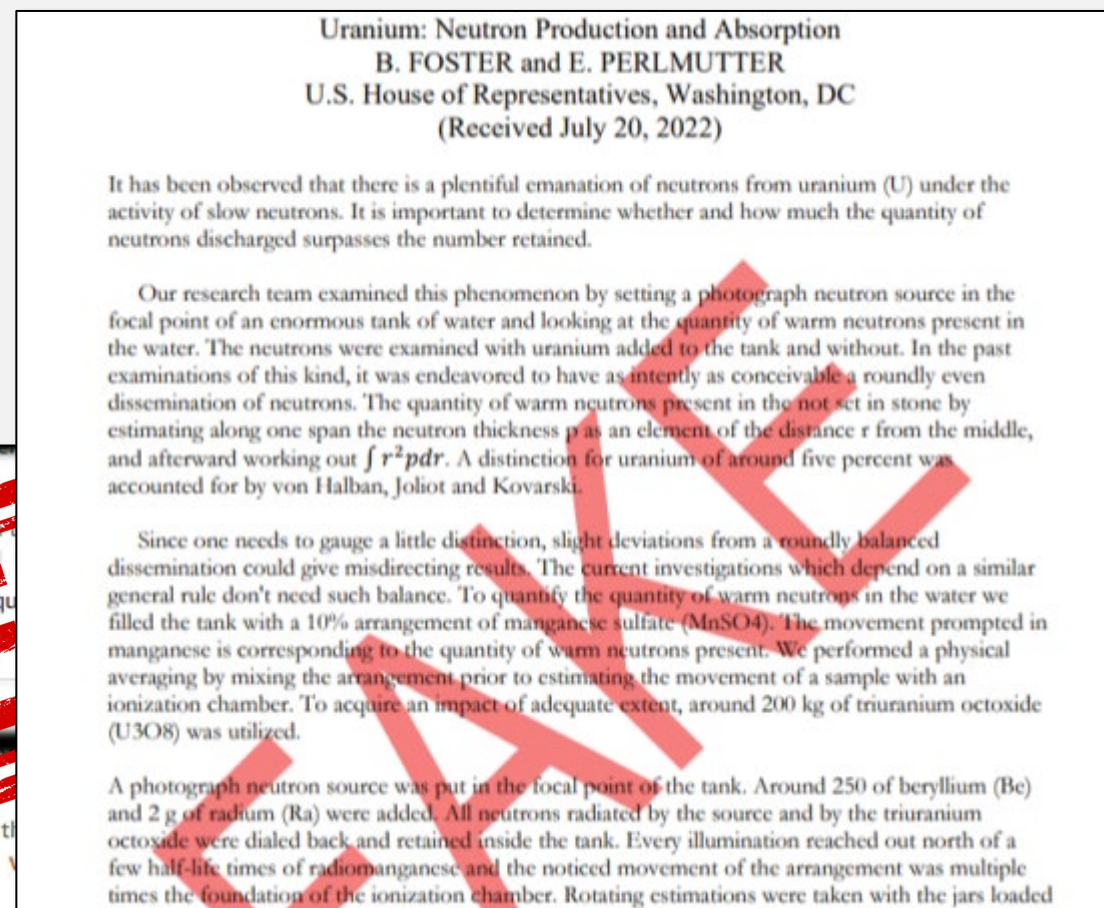
**News 2:** Maj Richard Scott, 40, is accused of driving at speeds of up to 95mph (153km/h) in bad weather before <u>the fatal crash that claimed the lives of two young children. The incident occurred on the A34 motorway near Newbury, Berkshire, last Saturday. Eyewitnesses reported seeing Maj Scott's car weaving in and out of traffic before losing control and colliding with another vehicle ...</u>

# Motivation

西 湖 大 學
WESTLAKE UNIVERSITY

**AI writings are hard to detect.**

- Linguistics experts identified AI-generated content **correctly only 38.9%** of the time.

- None of the **72 experts** correctly identified all four writing samples given to them.

Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. 2023. J. Elliott Casal and Matt Kessler. *Research Methods in Applied Linguistics*


西 湖 大 學
WESTLAKE UNIVERSITY


# Motivation

## Deepfake of Text Content

# Motivation

西 湖 大 學
WESTLAKE UNIVERSITY

**AIGC vs. Human Writing**

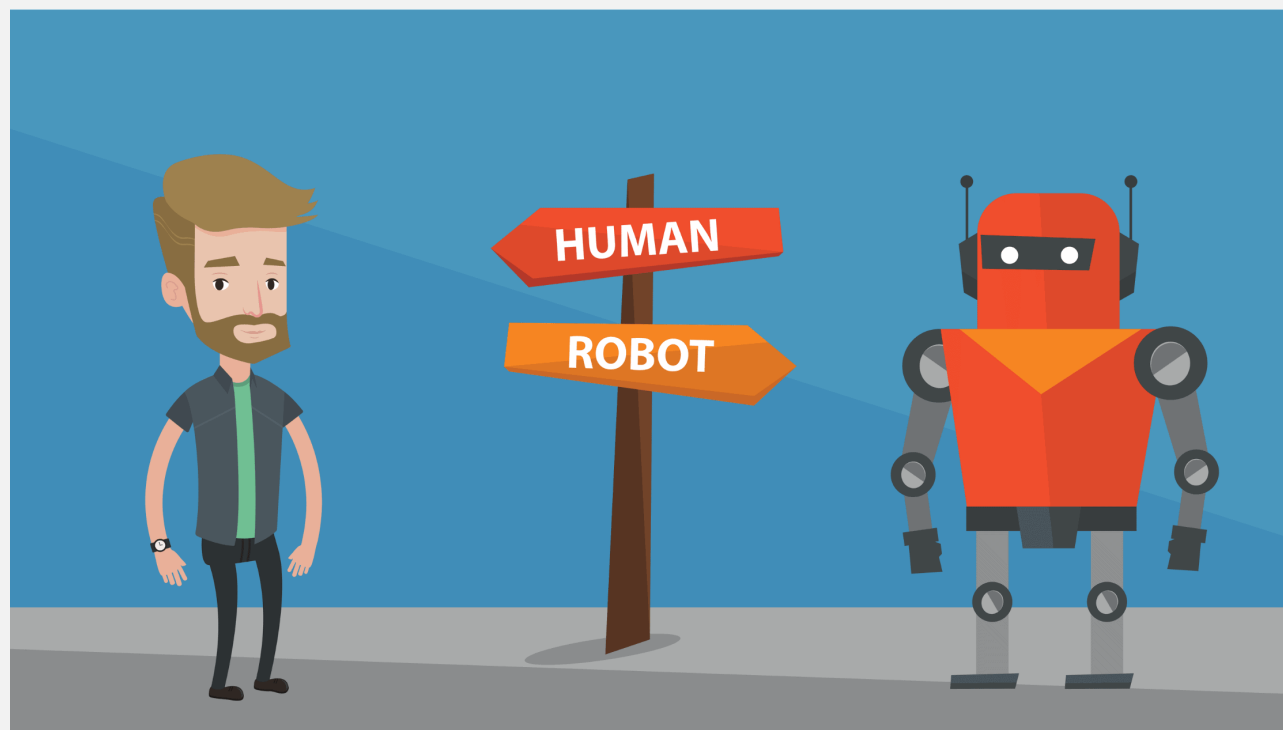
Trustworthy AI:

- Build trust and credibility in communication

- Avoid misuse of AI technology

Demand Tools:

- Machine-generated text detector

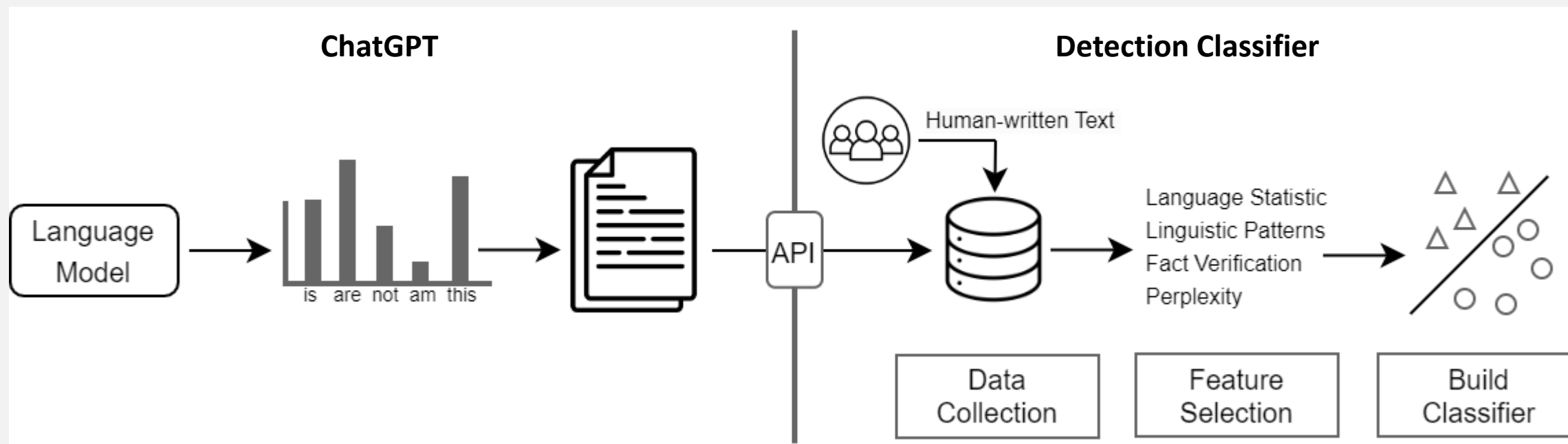
# Motivation

**Key Factors of Machine-Generated Text Detector**

- **Source models**:
  - ➢ What source models are supported by the detector?
  - ➢ Do we need samples for the source models for training?
  - ➢ Do we need access to the model parameters and architectures?

- **Languages**: what languages are supported by the detector?

- **Domains**: what domains (like news, review, and paper) are supported by the detector?

# Existing Approaches

## Approach I: Trained Detector

**For example:**

OpenAI AI classifier: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

CrossPlag AI content detector: https://crossplag.com/ai-content-detector

CopyLeaks AI content detector: https://copyleaks.com/ai-content-detector

# Existing Approaches

西 湖 大 學
WESTLAKE UNIVERSITY

**Approach I: Trained Detector**

Issues:

- **Model/domain/language-specific**: the trained detector overfits the training data, which generally covers specific source models, domains, and languages.

- **Limited dataset**: the collected datasets are relatively limited compared to large scale pretraining corpus.

- **High cost**: collecting LLM generations cost money, and training detectors with a big dataset costs computational resources.

# Existing Approaches

西 湖 大 學
WESTLAKE UNIVERSITY

**Approach II: Zero-Shot Detector**

Benefits:

- **No training**: using a pre-trained model for the detection.

- **Cover various domains and languages**: the pre-trained models are usually trained on large amounts of data covering various domains and languages.

# Existing Approaches

**Approach II: Zero-Shot Detector - Likelihood**

**Likelihood:** $$\frac{1}{N} \sum_j \log p(x_j | x_{<j})$$

**Intuition:** If a text has a very high likelihood according to a certain model, it might be an indication that the text was generated by that model or a similar one.

西 湖 大 學
WESTLAKE UNIVERSITY

# Existing Approaches

**Approach II: Zero-Shot Detector - DetectGPT**

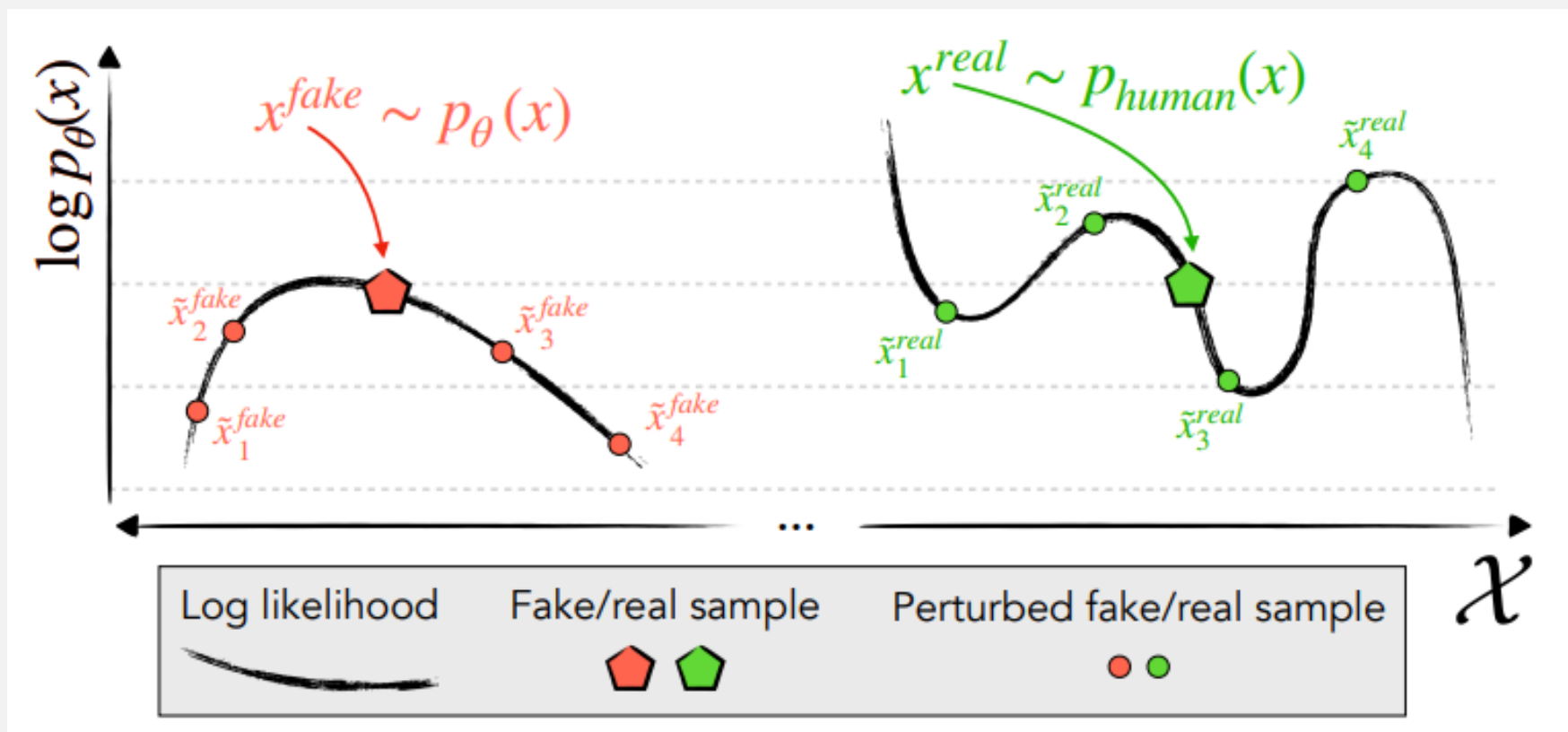
**DetectGPT:** $$\mathbf{d}(x, p_\theta, q_\varphi) = \log p_\theta(x) - \mathbb{E}_{\tilde{x} \sim q_\varphi(\cdot|x)}\left[\log p_\theta(\tilde{x})\right]$$

**Hypothesis**: Minor rewrites of model-generated text tend to have lower log probability under the model than the original sample, while minor rewrites of human-written text may have higher or lower log probability than the original sample.

DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. 2023.
Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, Chelsea Finn. *ICML*

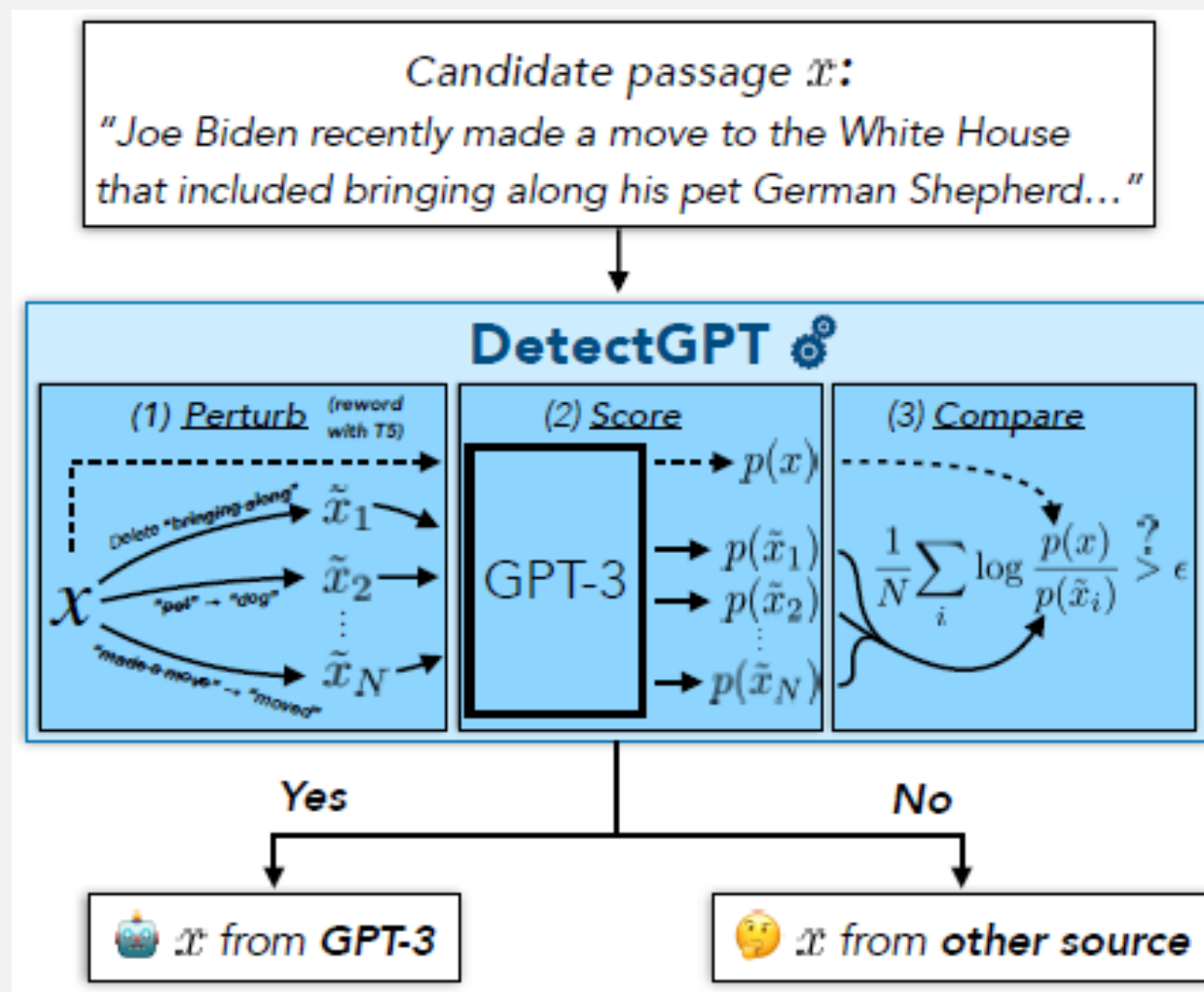
# Existing Approaches

**Approach II: Zero-Shot Detector - DetectGPT**

## DetectGPT - Probability Curvature

# Existing Approaches

**Approach II: Zero-Shot Detector - DetectGPT**

**Detection process:**

1) Perturb
2) Score
3) Compare

# Existing Approaches

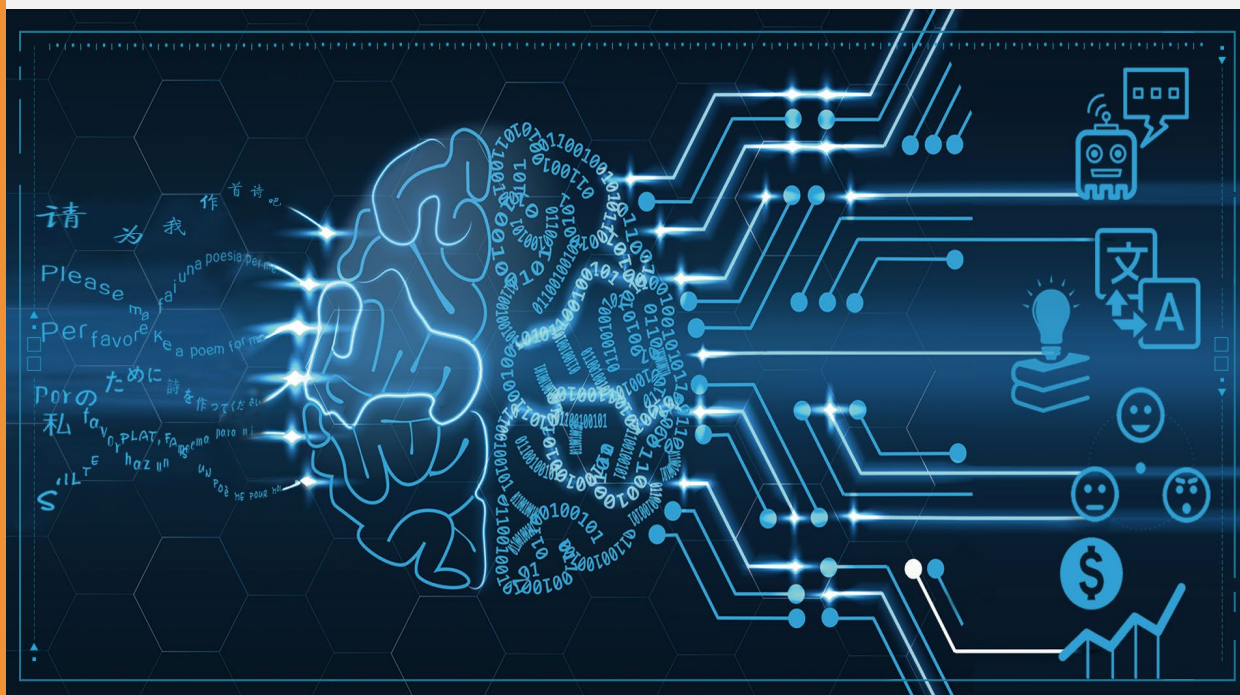
**Approach II: Zero-Shot Detector - DetectGPT**

Issues:

- **Model specific**: DetectGPT requires the source model to achieve accurate detection of the model-generated text.

- **Slow detection**: DetectGPT can provide good accuracy but slow execution because of multiple model/api calls.

- **Hight cost**: DetectGPT calls API service by hundred times to performance one detection, which costs a lot of computational resources or service budget.

# CONTENTS


01 Background

02 **Fast-DetectGPT**

03 Experiments

04 Code and Demo


# Fast-DetectGPT

**Key Features**

- **A zero-shot detector:** no training

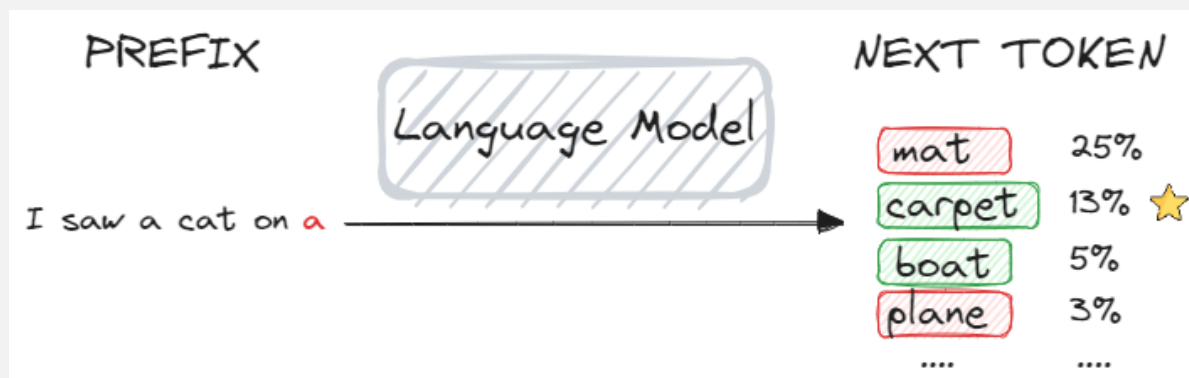
- **Use pretrained LLM:** robust across various domains and languages.

- **Use fixed LLMs to detect various LLMs**: robust across source models.

- **Use small LLMs to detect large LLMs**: reduce costs

# Fast-DetectGPT

西 湖 大 學
**WESTLAKE UNIVERSITY**

## Key Intuition

**Hypothesis**: Humans and machines tend to select different words during the text-generation process, with machines exhibiting a propensity for choosing words with higher model probabilities.

The hypothesis is **rooted in the fact** that LLMs, pre-trained on the largescale corpus, mirror **human collective writing behaviors** instead of **human individual writing behavior**, resulting in a discrepancy in their word choices given a context.

**Fast-DetectGPT**

西 湖 大 學
WESTLAKE UNIVERSITY

## Conditional Probability Curvature

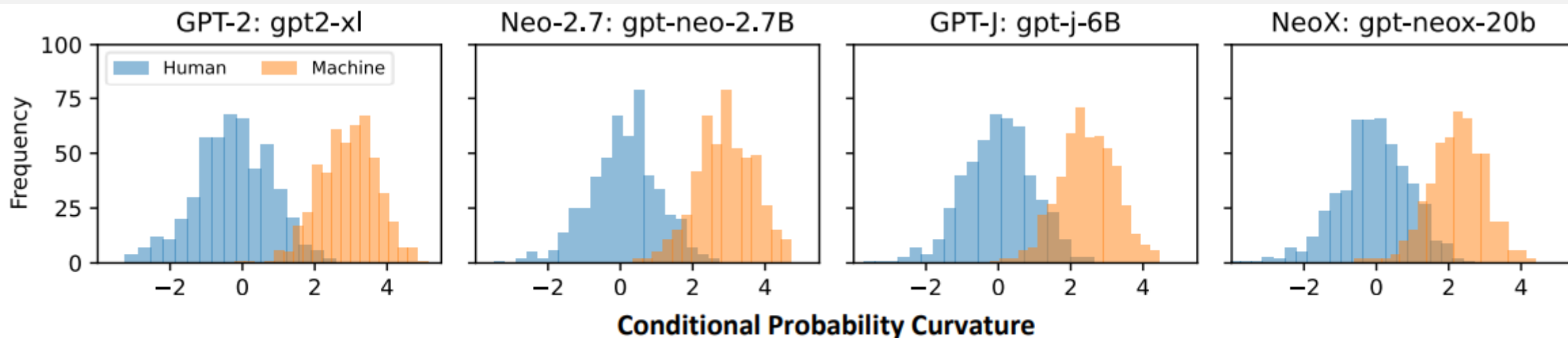



Figure 1: Distribution of *conditional probability curvatures* of the original human-written passages and the machine-generated passages by four source models on 30-token prefix from XSum.

# Fast-DetectGPT

## Conditional Probability Function

Good properties:

- Input is fixed for sampling and scoring

- One forward pass to obtain predictive distribution

- Conditional independent sampling

$$p_\theta(\tilde{x}|x) = \prod_j p_\theta(\tilde{x}_j|x_{<j})$$

.vs.

Probability function:

```
torch.distributions.categorical.Categorical(logits=lprobs).sample([10000])
```

$$= \prod_j p_\theta(\tilde{x}_j|\tilde{x}_{<j})$$

DetectGPT

# Fast-DetectGPT

西湖大學
WESTLAKE UNIVERSITY

## Conditional Probability Curvature

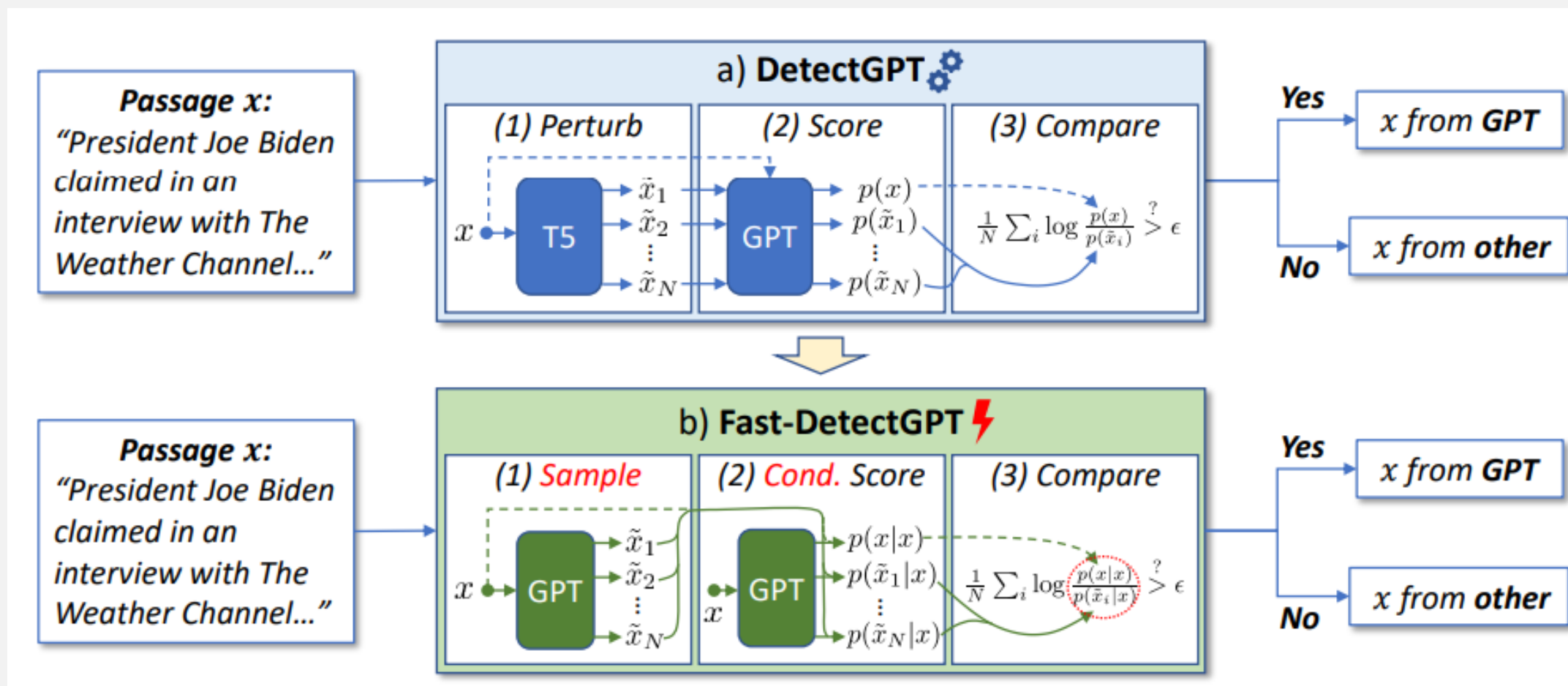
$$\mathbf{d}(x, p_\theta, q_\varphi) = \frac{\log p_\theta(x|x) - \tilde{\mu}}{\tilde{\sigma}}, \tag{3}$$

where

$$\tilde{\mu} = \mathbb{E}_{\tilde{x} \sim q_\varphi(\tilde{x}|x)} \left[ \log p_\theta(\tilde{x}|x) \right] \quad \text{and} \quad \tilde{\sigma}^2 = \mathbb{E}_{\tilde{x} \sim q_\varphi(\tilde{x}|x)} \left[ (\log p_\theta(\tilde{x}|x) - \tilde{\mu})^2 \right]. \tag{4}$$

vs.

$$\mathbf{d}(x, p_\theta, q_\varphi) = \log p_\theta(x) - \mathbb{E}_{\tilde{x} \sim q_\varphi(\cdot|x)} \left[ \log p_\theta(\tilde{x}) \right]$$

Probability Curvature used in DetectGPT

# Fast-DetectGPT

西湖大學 WESTLAKE UNIVERSITY

## Detection Process

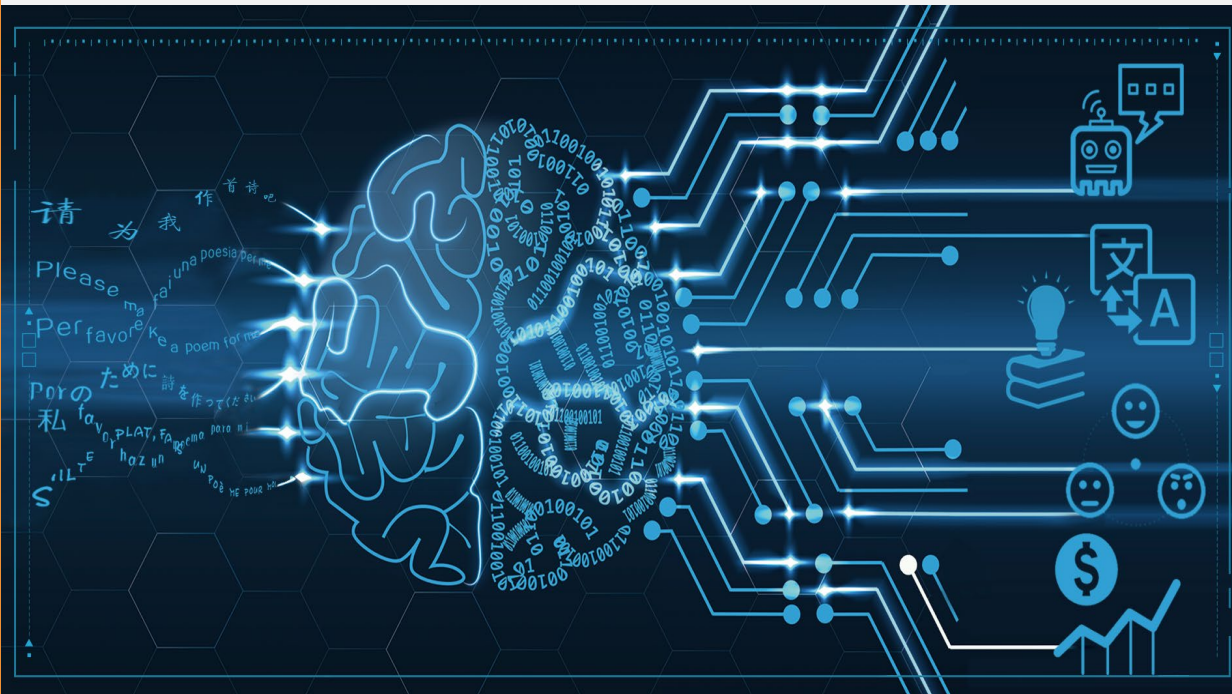
# CONTENTS


01 Background

02 Fast-DetectGPT

03 **Experiments**

04 Code and Demo


# Experiments

**Task Settings**

- **The White-Box Setting:**

We have the privilege of accessing the possible source model. We use the source model to aid in scoring the candidate passage to inform the classification decision in the setting.

- **The Black-Box Setting:**

We operate without access to the source model. Instead, we rely on surrogate models to score the passage.

# Experiments

## Settings – Source Models

| Model | Model File/Service | Parameters | Training Corpus |
| --- | --- | --- | --- |
| mGPT (Shliazhko et al., 2022) | sberbank-ai/mGPT | 1.3B | Wikipedia and Colossal Clean Crawled corpus. |
| GPT-2 (Radford et al., 2019) | gpt2-xl | 1.5B | English WebText without Wikipedia. |
| PubMedGPT (Bolton et al., 2022) | stanford-crfm/pubmedgpt | 2.7B | Biomedical abstracts and papers from the Pile. |
| OPT-2.7 (Zhang et al., 2022) | facebook/opt-2.7b | 2.7B | A larger dataset including the Pile. |
| Neo-2.7 (Black et al., 2021) | EleutherAI/gpt-neo-2.7B | 2.7B | The Pile (Gao et al., 2020). |
| GPT-J (Wang & Komatsuzaki, 2021) | EleutherAI/gpt-j-6B | 6B | The Pile (Gao et al., 2020). |
| BLOOM-7.1 (Scao et al., 2022) | bigscience/bloom-7b1 | 7.1B | ROOTS corpus (Laurençon et al., 2022). |
| OPT-13 (Zhang et al., 2022) | facebook/opt-13b | 13B | A larger dataset including the Pile. |
| Llama-13 (Touvron et al., 2023a) | huggyllama/llama-13b | 13B | CommonCrawl, C4, Github, Wikipedia, Books, ... |
| Llama2-13 (Touvron et al., 2023b) | TheBloke/Llama-2-13B-fp16 | 13B | A new mix of publicly available online data. |
| NeoX (Black et al., 2022) | EleutherAI/gpt-neox-20b | 20B | The Pile (Gao et al., 2020). |
| GPT-3 (Brown et al., 2020) | OpenAI/davinci | 175B | CommonCrawl, WebText, English Wikipedia, ... |
| ChatGPT (OpenAI, 2022) | OpenAI/gpt-3.5-turbo | 175B | CommonCrawl, WebText, English Wikipedia, ... |
| GPT-4 (OpenAI, 2023) | OpenAI/gpt-4 | NA | NA |

**Experiments** 西 湖 大 學 WESTLAKE UNIVERSITY

**Settings – Datasets**

- **Datasets**:
    - XSum (news), WritingPrompt (story), SQuAD (wikipedia), PubMedQA (technical)
    - WMT16 English and German

1:1 positive and negative samples, where a machine-generated text is generated by the source model using the first 30 tokens of human-written text as the prefix.

# Experiments

**Overview**

| Method | 5-Model Generations ↑ | ChatGPT/GPT-4 Generations ↑ | Speedup ↑ |
|---|---|---|---|
| DetectGPT | 0.9554 | 0.7225 | 1x |
| Fast-DetectGPT | **0.9887** (relative↑ 74.7%) | **0.9338** (relative↑ 76.1%) | **340x** |

Table 1: Dete... machine-generated text... applied to the methods d... the *black-box setting* (ut... s are averaged from dat... where the 'relative↑' is calculated by $(new - old)/(1.0 - old)$, representing how much improvement has been made relative to the maximum possible improvement. Speedup assessments were conducted using the XSum news dataset, with computations on a Tesla A100 GPU.

> DetectGPT uses GPU batch processing, splitting 100 perturbations into 10 batches. DetectGPT costs about **22 hours** in five runs against five open-source models, while Fast-DetectGPT only costs **4 minutes**.

<cut/>**Experiments**

## 5-Model Generations – The White-Box Setting

| Method | GPT-2 | OPT-2.7 | Neo-2.7 | GPT-J | NeoX | Avg. |
|---|---|---|---|---|---|---|
| **The White-Box Setting** | | | | | | |
| Likelihood | 0.9125 | 0.8963 | 0.8900 | 0.8480 | 0.7946 | 0.8683 |
| Entropy | 0.5174 | 0.4830 | 0.4898 | 0.5005 | 0.5333 | 0.5048 |
| LogRank | 0.9385 | 0.9223 | 0.9226 | 0.8818 | 0.8313 | 0.8993 |
| LRR | 0.9601 | 0.9401 | 0.9522 | 0.9179 | 0.8793 | 0.9299 |
| DNA-GPT ◇ | 0.9024 | 0.8797 | 0.869 | 0.8227 | 0.7826 | 0.8513 |
| NPR ◇ | 0.9948† | 0.9832† | 0.9883 | 0.9500 | 0.9065 | 0.9645 |
| DetectGPT (T5-3B/*) ◇ | 0.9917 | 0.9758 | 0.9797 | 0.9353 | 0.8943 | 0.9554 |
| Fast-DetectGPT (*/*) | **0.9967** | **0.9908** | 0.9940† | **0.9866** | **0.9754** | **0.9887** |
| (Relative↑) | 60.2% | 62.0% | 70.4% | 79.3% | 76.7% | 74.7% |
| **The Black-Box Setting** | | | | | | |
| DetectGPT (T5-3B/Neo-2.7) ◇ | 0.8517 | 0.8390 | 0.9797 | 0.8575 | 0.8400 | 0.8736 |
| Fast-DetectGPT (GPT-J/Neo-2.7) | 0.9834 | 0.9572 | **0.9984** | 0.9592† | 0.9404† | 0.9677† |
| (Relative↑) | 88.8% | 73.4% | 92.1% | 71.4% | 62.8% | 74.5% |

# Experiments

**5-Model Generations – The Black-Box Setting**

| Method | GPT-2 | OPT-2.7 | Neo-2.7 | GPT-J | NeoX | Avg. |
|---|---|---|---|---|---|---|
| **The White-Box Setting** | | | | | | |
| Likelihood | 0.9125 | 0.8963 | 0.8900 | 0.8480 | 0.7946 | 0.8683 |
| Entropy | 0.5174 | 0.4830 | 0.4898 | 0.5005 | 0.5333 | 0.5048 |
| LogRank | 0.9385 | 0.9223 | 0.9226 | 0.8818 | 0.8313 | 0.8993 |
| LRR | 0.9601 | 0.9401 | 0.9522 | 0.9179 | 0.8793 | 0.9299 |
| DNA-GPT ◇ | 0.9024 | 0.8797 | 0.869 | 0.8227 | 0.7826 | 0.8513 |
| NPR ◇ | 0.9948† | 0.9832† | 0.9883 | 0.9500 | 0.9065 | 0.9645 |
| DetectGPT (T5-3B/*) ◇ | 0.9917 | 0.9758 | 0.9797 | 0.9353 | 0.8943 | 0.9554 |
| Fast-DetectGPT (*/*) | **0.9967** | **0.9908** | 0.9940† | **0.9866** | **0.9754** | **0.9887** |
| *(Relative↑)* | *60.2%* | *62.0%* | *70.4%* | *79.3%* | *76.7%* | *74.7%* |
| **The Black-Box Setting** | | | | | | |
| DetectGPT (T5-3B/Neo-2.7) ◇ | 0.8517 | 0.8390 | 0.9797 | 0.8575 | 0.8400 | 0.8736 |
| Fast-DetectGPT (GPT-J/Neo-2.7) | 0.9834 | 0.9572 | **0.9984** | 0.9592† | 0.9404† | 0.9677† |
| *(Relative↑)* | *88.8%* | *73.4%* | *92.1%* | *71.4%* | *62.8%* | *74.5%* |

# Experiments

## ChatGPT / GPT-4 Generations – The Black-Box Setting

| Method | ChatGPT | | | | GPT-4 | | | |
|---|---|---|---|---|---|---|---|---|
| | XSum | Writing | PubMed | Avg. | XSum | Writing | PubMed | Avg. |
| RoBERTa-base | 0.9150 | 0.7084 | 0.6188 | 0.7474 | 0.6778 | 0.5068 | 0.5309 | 0.5718 |
| RoBERTa-large | 0.8507 | 0.5480 | 0.6731 | 0.6906 | 0.6879 | 0.3821 | 0.6067 | 0.5589 |
| GPTZero | **0.9952** | 0.9292 | 0.8799 | 0.9348 | **0.9815** | 0.8262 | 0.8482 | 0.8853 |
| Likelihood (Neo-2.7) | 0.9578 | 0.9740 | 0.8775 | 0.9364 | 0.7980 | 0.8553 | 0.8104 | 0.8212 |
| Entropy (Neo-2.7) | 0.3305 | 0.1902 | 0.2767 | 0.2658 | 0.4360 | 0.3702 | 0.3295 | 0.3786 |
| LogRank(Neo-2.7) | 0.9582 | 0.9656 | 0.8687 | 0.9308 | 0.7975 | 0.8286 | 0.8003 | 0.8088 |
| LRR (Neo-2.7) | 0.9162 | 0.8958 | 0.7433 | 0.8518 | 0.7447 | 0.7028 | 0.6814 | 0.7096 |
| DNA-GPT (Neo-2.7) | 0.9124 | 0.9425 | 0.7959 | 0.8836 | 0.7347 | 0.8032 | 0.7565 | 0.7648 |
| NPR (T5-11B/Neo-2.7) | 0.7899 | 0.8924 | 0.6784 | 0.7869 | 0.5280 | 0.6122 | 0.6328 | 0.5910 |
| DetectGPT (T5-11B/Neo-2.7) | 0.8416 | 0.8811 | 0.7444 | 0.8223 | 0.5660 | 0.6217 | 0.6805 | 0.6228 |
| Fast-Detect (GPT-J/Neo-2.7) | 0.9907 | **0.9916** | **0.9021** | **0.9615** | 0.9067 | **0.9612** | **0.8503** | **0.9061** |
| *(Relative ↑)* | *94.1%* | *92.9%* | *61.7%* | *78.3%* | *78.5%* | *89.7%* | *53.1%* | *75.1%* |

\* Only the black-box setting

**Experiments**

## GPT-3 Generations – The Black-Box and White-Box Setting

| Method | XSum | WritingPrompts | PubMedQA | Avg. |
|---|---|---|---|---|
| RoBERTa-base | 0.8986 | 0.9282 | 0.6370 | 0.8212 |
| RoBERTa-large | 0.9325 | 0.9113 | 0.6894 | 0.8444 |
| GPTZero | 0.4860 | 0.6009 | 0.4246 | 0.5038 |
| Likelihood (GPT-3) ◇ | 0.76 | 0.87 | 0.64 | 0.76 |
| DetectGPT (T5-11B/GPT-3) ◇ | 0.84 | 0.87 | **0.84** | 0.85 |
| Likelihood (Neo-2.7) | 0.8307 | 0.8496 | 0.5668 | 0.7490 |
| Entropy (Neo-2.7) | 0.3923 | 0.3049 | 0.5358 | 0.4110 |
| LogRank(Neo-2.7) | 0.8096 | 0.8320 | 0.5527 | 0.7314 |
| LRR (Neo-2.7) | 0.6687 | 0.7410 | 0.4917 | 0.6338 |
| DNA-GPT (Neo-2.7) | 0.8209 | 0.8354 | 0.5761 | 0.7441 |
| NPR (T5-11B/Neo-2.7) | 0.8032 | 0.7847 | 0.6342 | 0.7407 |
| DetectGPT (T5-11B/GPT-2) | 0.8043 | 0.7699 | 0.6915 | 0.7552 |
| DetectGPT (T5-11B/Neo-2.7) | 0.8455 | 0.7818 | 0.6977 | 0.7750 |
| DetectGPT (T5-11B/GPT-J) | 0.8261 | 0.7666 | 0.6644 | 0.7524 |
| Fast-DetectGPT (GPT-J/GPT-2) | 0.9137 | 0.9533* | 0.7589* | **0.8753** |
| Fast-DetectGPT (GPT-J/Neo-2.7) | **0.9396** | 0.9492 | 0.7225 | 0.8704* |
| Fast-DetectGPT (GPT-J/GPT-J) | 0.9329* | **0.9568** | 0.6664 | 0.8520 |

# **Usability Analysis**

西湖大學
WESTLAKE UNIVERSITY

## **Low False Alarm, High Recall**

False alarm: 1%

Recall on ChatGPT:

- Fast-Detect: **87%**

- DetectGPT: **6%**

Recall on GPT-4:

- Fast-Detect: **44%**

- DetectGPT: **0%**

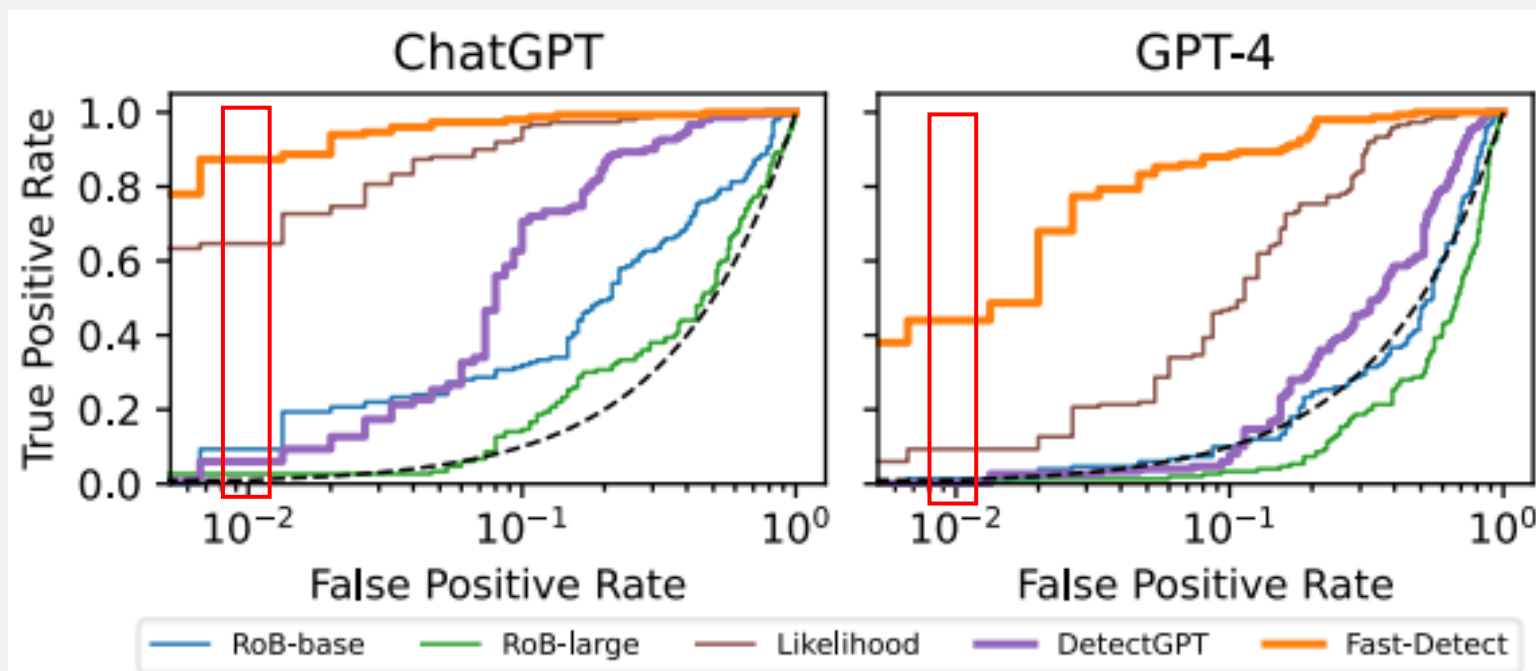



Figure 3: ROC curves in log scale evaluated on stories (WritingPrompts), where the dash lines denote the random classifier.

# Usability Analysis

**Robustness across Domains and Languages**

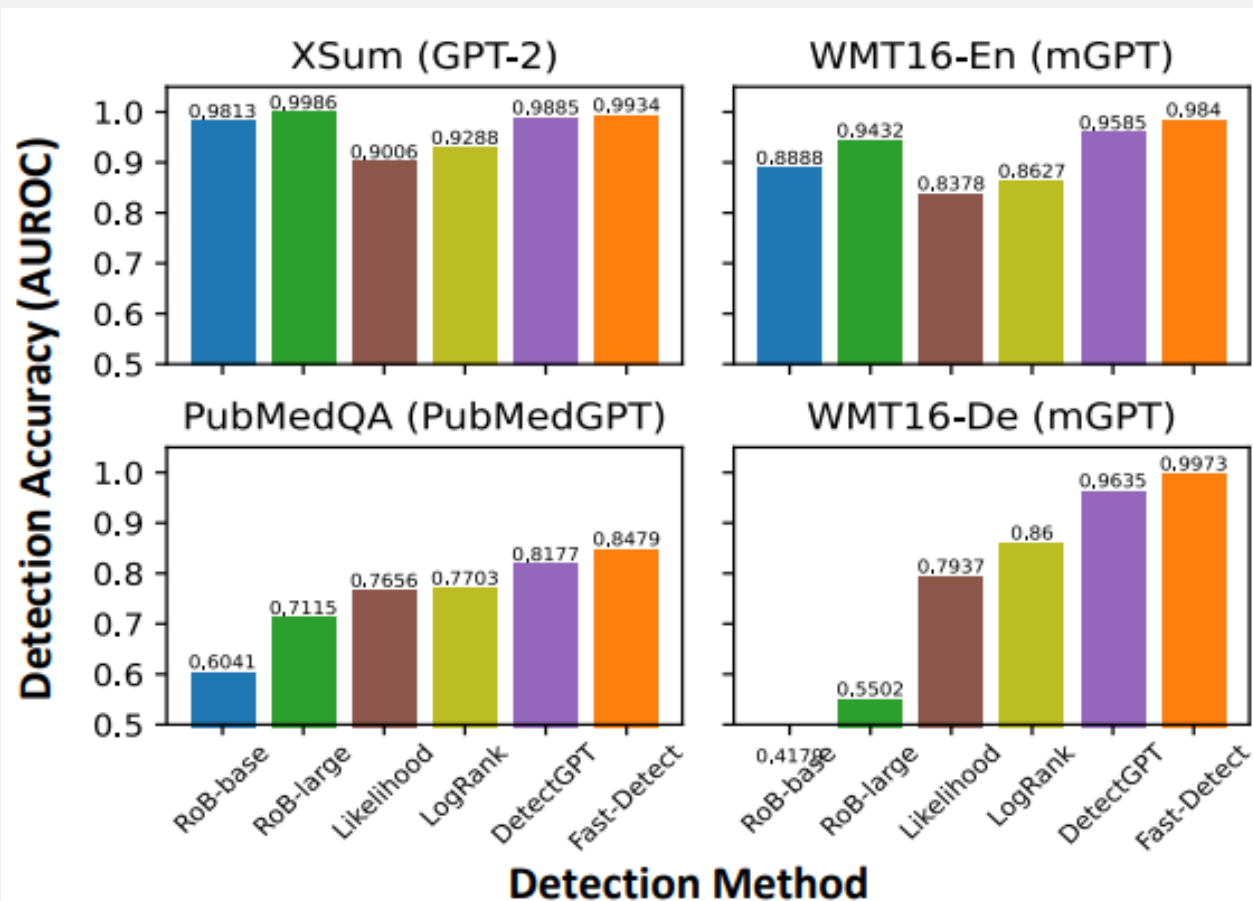



Figure 5: Compare with supervised detectors, evaluated in AUROC. We generate 200 test samples for each dataset and source model.

# Usability Analysis

**Robustness on Different Passage Lengths**

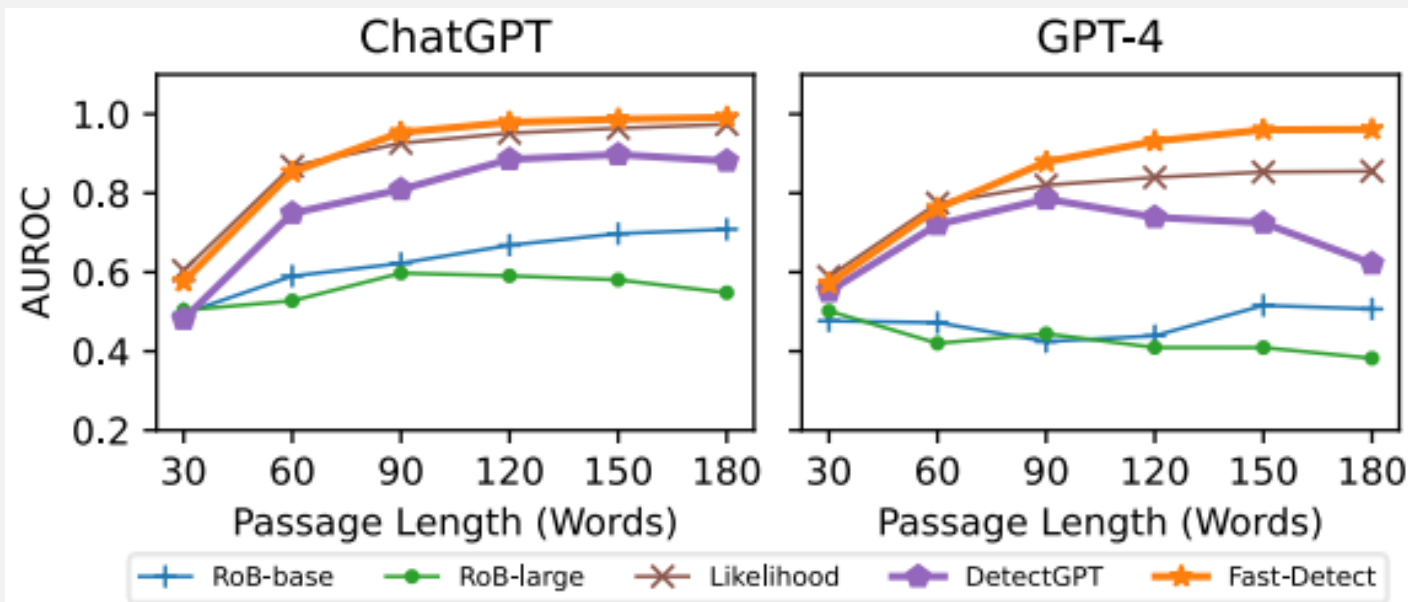



Figure 4: Detection accuracy (AUROC) on story passages (WritingPrompts) truncated to target number of words.

# CONTENTS

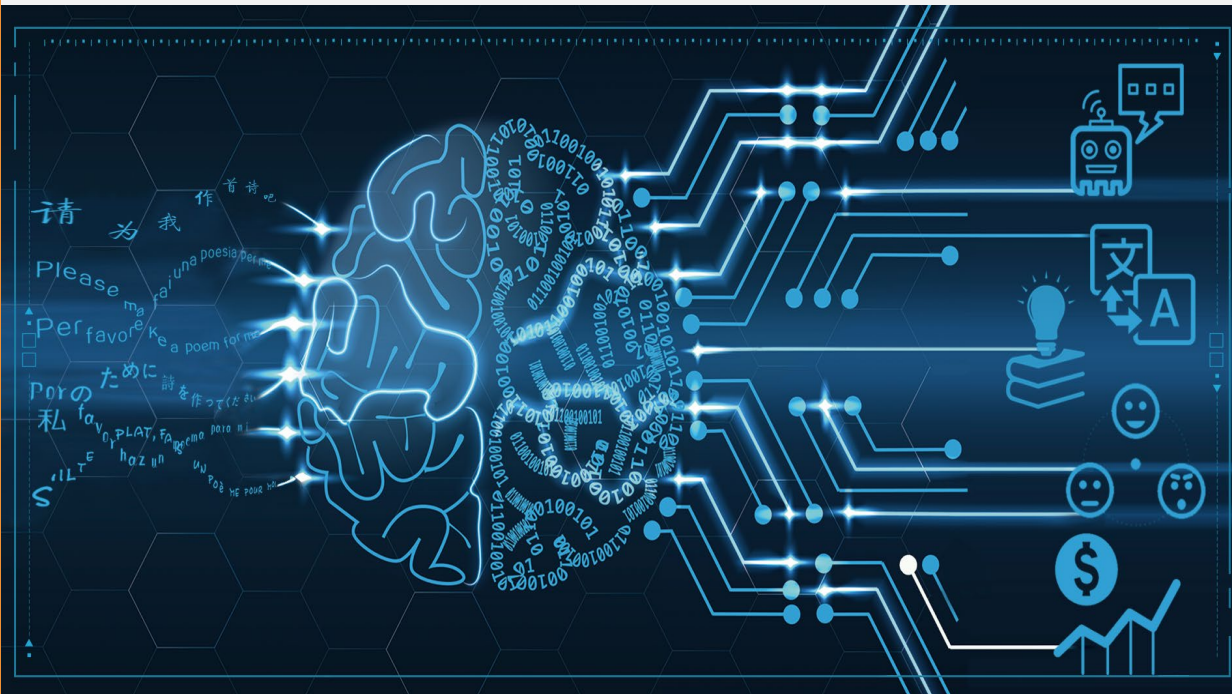

01    Background

02    Fast-DetectGPT

03    Experiments

04    **Code and Demo**


# Github

https://github.com/bao
guangsheng/fast-detect-
gpt

| | | |
|---|---|---|
| 📁 exp_gpt3to4/data | remove pii and fix path. | 5 months ago |
| 📁 exp_main/data | generated data for main exper... | 5 months ago |
| 📁 local_infer_ref | reference distribution for local ... | 5 months ago |
| 📁 scripts | Update local_infer.py | 3 months ago |
| 📄 LICENSE | Initial commit | 8 months ago |
| 📄 README.md | Update README.md | 3 weeks ago |
| 📄 attack.sh | entry scripts for experiments | 5 months ago |
| 📄 gpt3to4.sh | add additional baselines | 4 months ago |
| 📄 main.sh | add additional baselines | 4 months ago |
| 📄 main_ext.sh | entry scripts for experiments | 5 months ago |

View all files

# Local Demo

西 湖 大 學
**WESTLAKE UNIVERSITY**

**python scripts/local_infer.py**

*Please enter your text: (Press Enter twice to start processing)*

**Input passage**

*Disguised as police, they broke through a fence on Monday evening and broke into the cargo of a Swiss-bound plane to take the valuable items. The audacious heist occurred at an airport in a small European country, leaving authorities baffled and airline officials in shock.*

**Output result**

*Fast-DetectGPT criterion is 1.9299, suggesting that the text has **a probability of 87% to be fake**.*

# Recap

## Deepfake of Text Content

**Recap** 西 湖 大 學

**AIGC vs. Human Writing**

Trustworthy AI:

- Build trust and credibility in communication

- Avoid misuse of AI technology

Demand Tools:

- Machine-generated text detector

**Recap**

**Zero-Shot Detector - DetectGPT**

Issues:

- **Model specific**: DetectGPT requires the source model to achieve accurate detection of the model-generated text.

- **Slow detection**: DetectGPT can provide good accuracy but slow execution because of multiple model/api calls.

- **Hight cost**: DetectGPT calls API service by hundred times to performance one detection, which costs a lot of computational resources or service budget.

# Recap

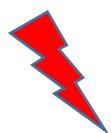
**Fast-DetectGPT:**

- **A zero-shot detector:** no training

- **Use pretrained LLM:** robust across various domains and languages.

- **Use fixed LLMs to detect various LLMs**: robust across source models.

- **Use small LLMs to detect large LLMs**: reduce costs

# Key Achievements

**Fast-DetectGPT**                    **DetectGPT**

LM: **6B** + **2.7B**                 LM: **11B** + **2.7B**

Cost **1** LM call                    Cost **100** LM calls

Speed **340x**                        Speed **1x**

Accuracy **96%**                      Accuracy **82%**

On ChatGPT generations

Additionally, Fast-DetectGPT **outperforms** commercial **GPTZero**.

Q & A

西 湖 大 學
WESTLAKE UNIVERSITY

# Thank you!

## Connect and further communicate.

知乎：西湖老博士

Twitter：@gshbao