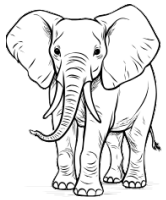


ACL 2024

Bangkok, Thailand

Website; Q&A

Watermarking for Large Language Models

Part IV: Post-hoc Detection

Xuandong Zhao

UC Berkeley

Yu-Xiang Wang

UC San Diego

Lei Li

CMU

Can we detect LLM-generated text directly? (without modifying model or decoding)

How to do post-hoc detection?

- Trained Classifiers

  Bag-of-words classifier e.g. Solaiman et al., 2019

  LLM classifier e.g. Zellers et al., 2019

- Zero-shot Classifiers

  Statistical Outlier Detection e.g. Techniques based on entropy, perplexity, n-gram frequencies

  DetectGPT by Mitchell et al. (2023)

  Theoretical limits of detectability (Sadasivan et al., 2023)

  Fast-DetectGPT by Bao et al. (2023)

OpenAI AI Classifier

January 31, 2023

New AI classifier for indicating AI-written text

We're launching a classifier trained to distinguish between AI-written and human-written text.

OpenAI AI Classifier: Training

- Fine-Tuning: Trained on a dataset consisting of human-written and AI-written text pairs on the same topics.
- Human Text Sources: Includes pretraining data and human demonstrations on InstructGPT prompts.
- AI Text Sources: Responses generated from various language models.
- Threshold Adjustment: The threshold is set to maintain a low false positive rate (FPR) in the web app. Text is marked as likely AI-written only if the classifier is highly confident.

OpenAI AI Classifier: Limitations

- **Text Length**: The classifier is very unreliable for short texts (below 1,000 characters).
- **False Positives**: Sometimes, human-written text will be incorrectly but confidently labeled as AI-written.
- **Lang**[illegible] than[illegible]
- **Prec**[illegible] cannot be reliably identified.
- **Evasion**: AI-written text can be edited to evade the classifier.
- **OOD**: Poorly calibrated for data outside of its training set.

As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy.

DetectGPT

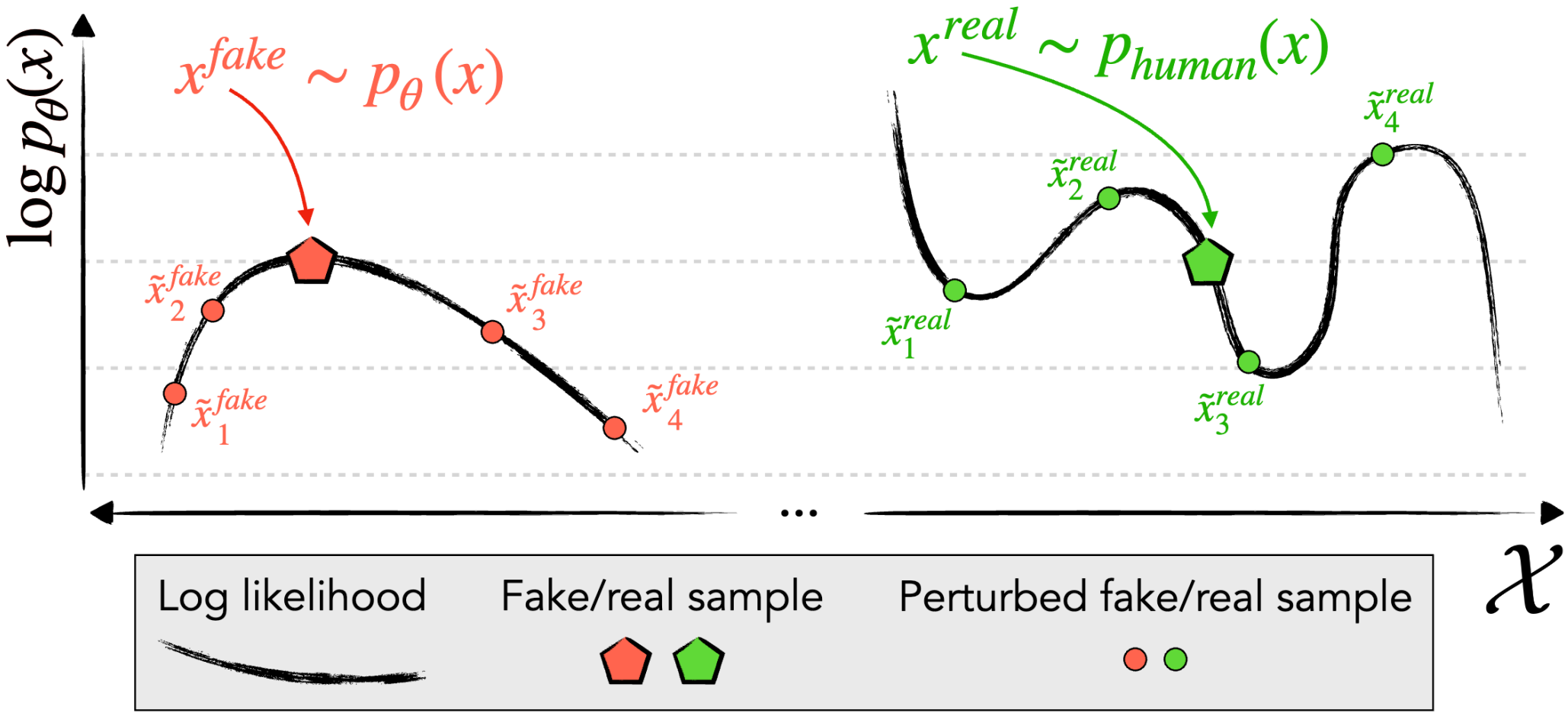
- Use the source LLM itself to detect its generations--"zero-shot”
- Idea:

  Use the structure of the log probability function around a given passage
- Hypothesis:

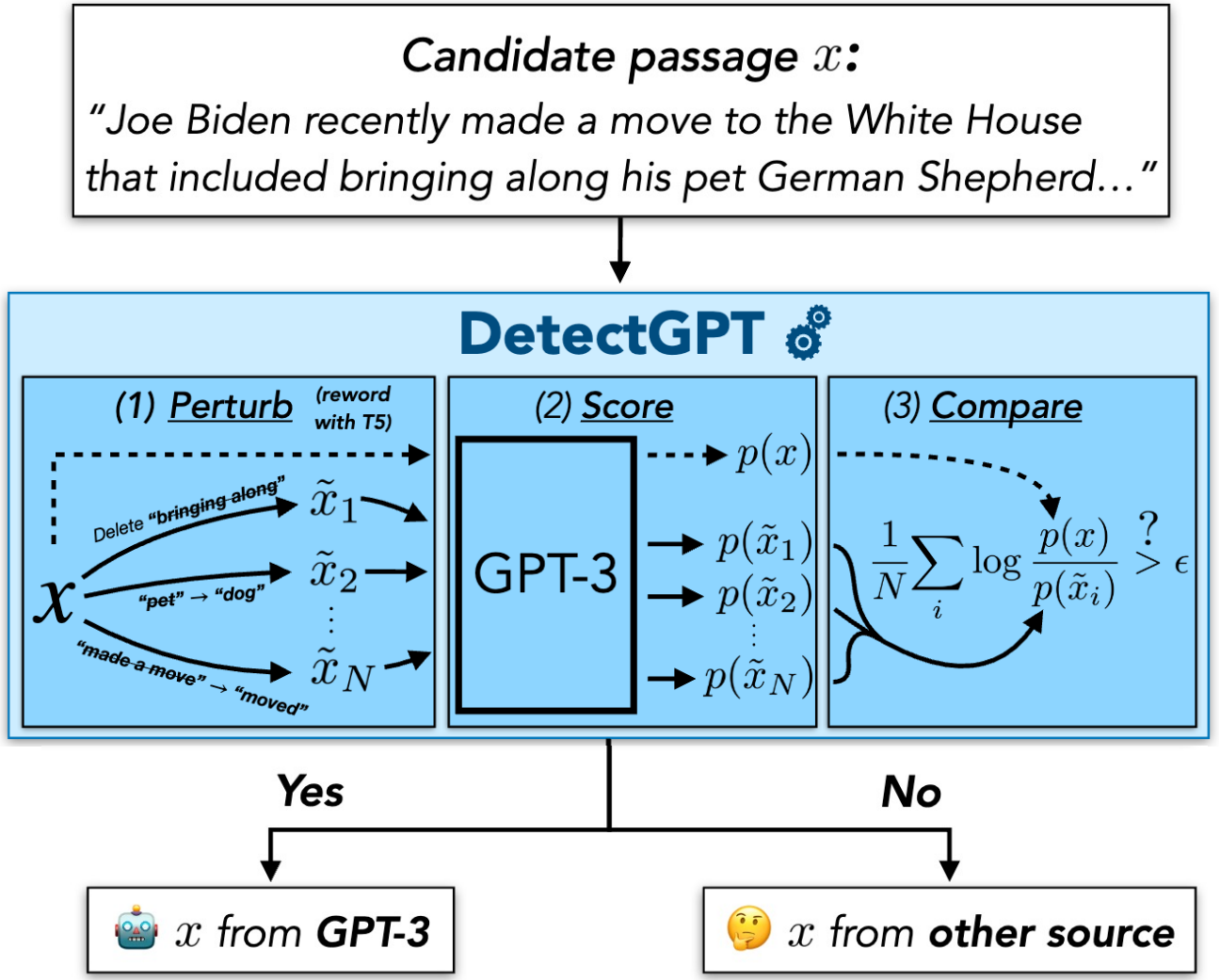
  Model samples lie near local maxima of the model's log probability function

  "If we slightly perturb model-generated text, the log probability tends to drop" (i.e., rephrase)

DetectGPT: Hypothesis about local structure of log probability



DetectGPT: Detection with Probability Curvature



Fast-DetectGPT: Conditional Probability Curvature

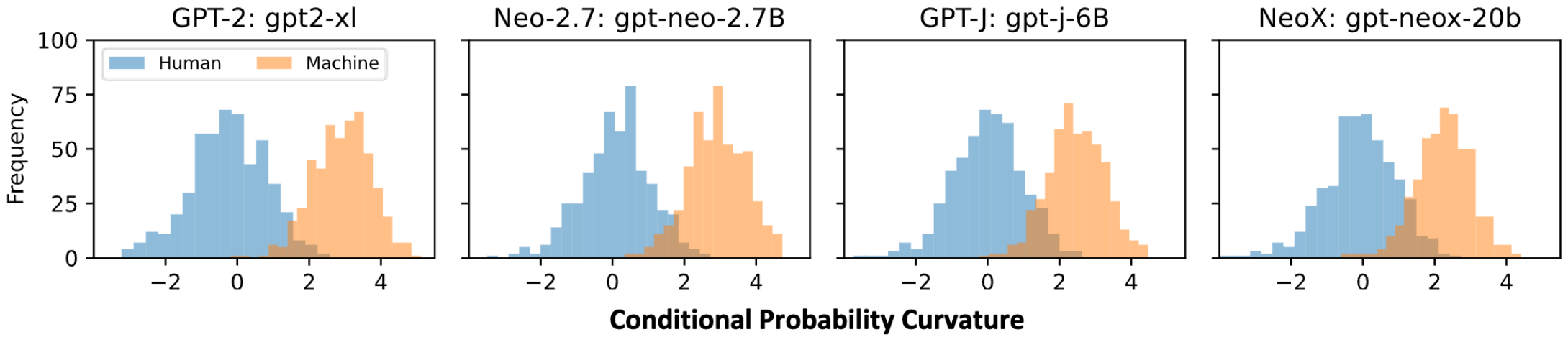
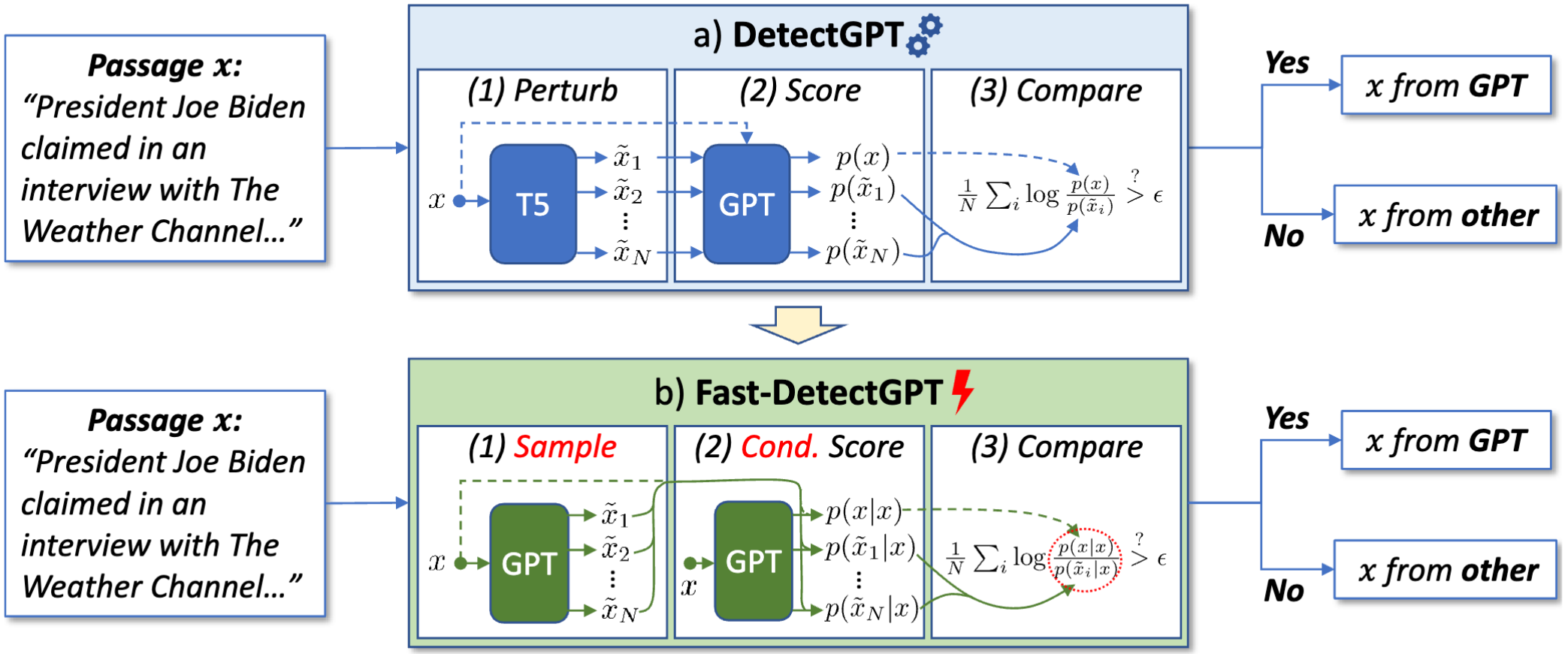


Figure 1: Distribution of *conditional probability curvatures* of the original human-written passages and the machine-generated passages by four source models on 30-token prefix from XSum.

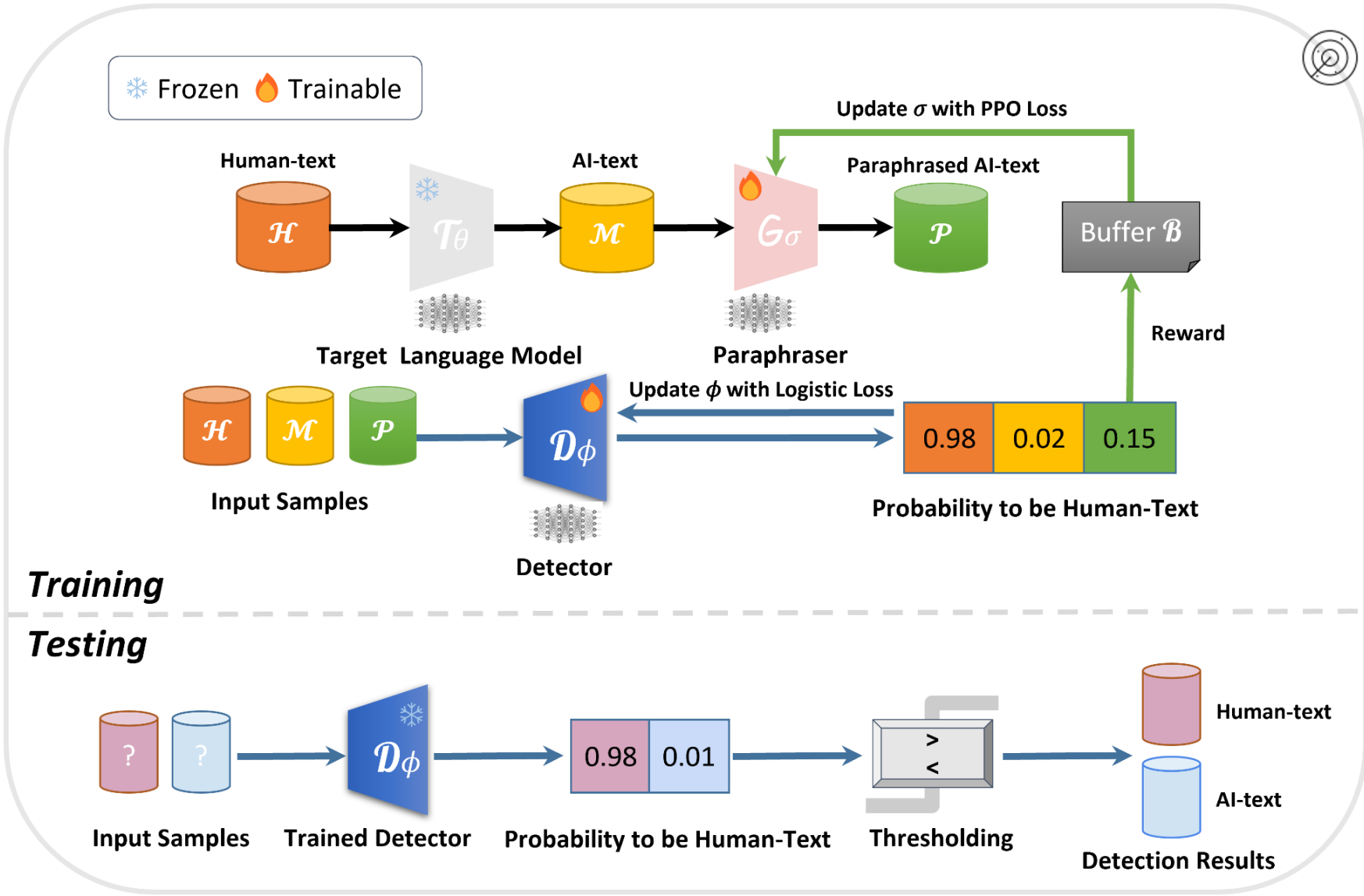
Fast-DetectGPT v.s. DetectGPT



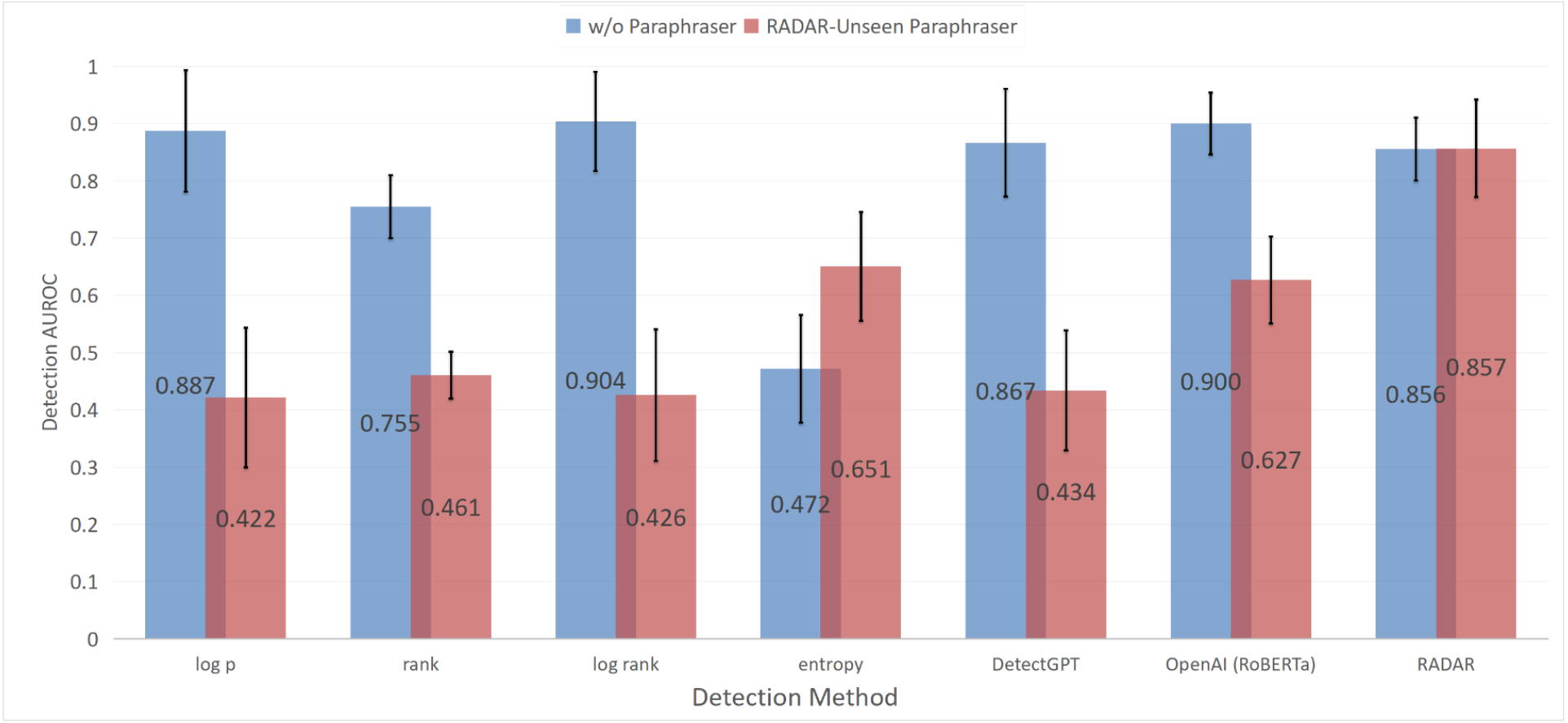
Results: AUROC and Speedup

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**

RADAR: Robust AI-Text Detection via Adversarial



RADAR Results



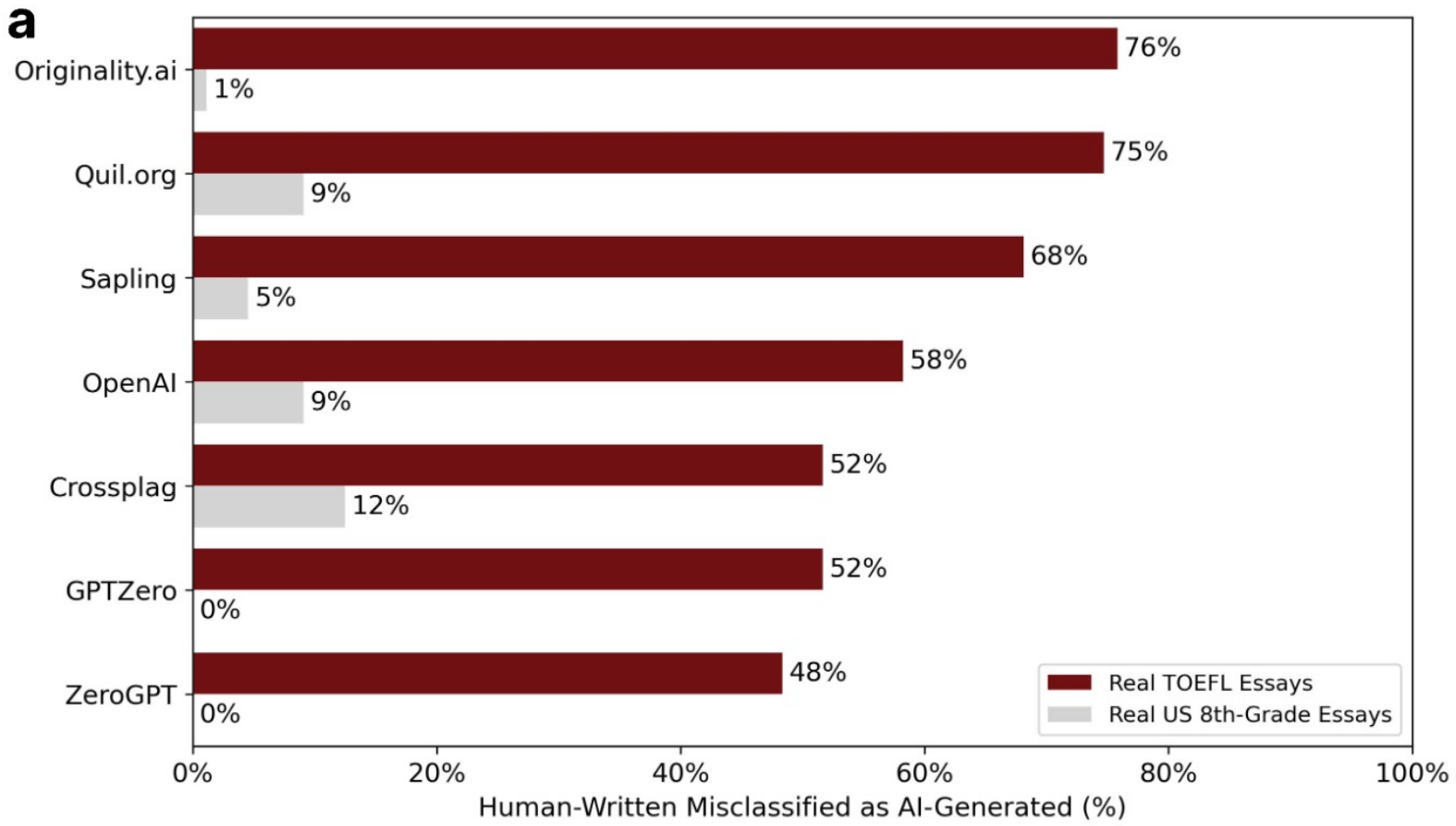
Are the trained and zero-shot classifiers robust?

Generative Model	Dataset	Unattacked	Dipper Paraphrasing	Query-free Substitution	Query-based Substitution
GPT-2-XL	XSum	84.4	35.2	25.9	**3.9**
	ELI5	70.6	36.7	21.2	**3.8**
ChatGPT	XSum	56.0	34.6	25.6	**4.5**
	ELI5	55.0	39.5	12.2	**6.5**
LLaMA-65B	XSum	59.3	49.0	25.5	**9.9**
	ELI5	60.5	53.1	31.4	**18.6**

Table 4: AUROC scores (%) of DetectGPT under various attack settings.

Shi et al. 2023 Red Teaming Language Model Detectors with Language Models

Are the trained and zero-shot classifiers robust?



Liang et al. 2023 GPT detectors are biased against non-native English writers