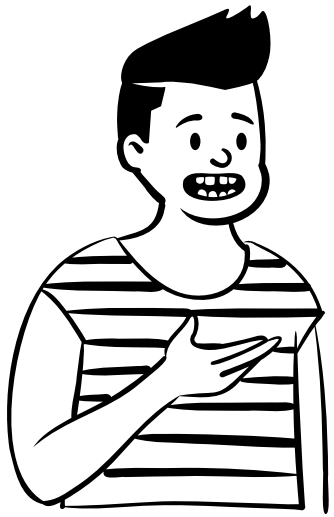
MGTBench: Benchmarking Machine-Generated Text Detection

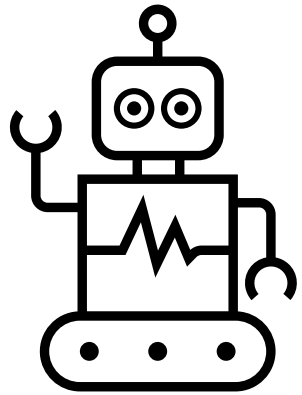
By Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, Yang Zhang

Presented by Maurice Shih

Introduction

?





MGTBench Overview

- First framework to benchmark machine generated text (MGT) detection methods
- Compose of three modules:
  - **Input Module** – Pre-processing on datasets (optimized for datasets from HuggingFace)
  - **Detection Module** – Applies different metric or model-based detections
  - **Evaluation Module** – Create evaluation metrics based on detection results

Metric-based Methods vs Model-based methods

- Metric-based methods - Use pre-trained LLMs to extract features from text
- Model-based methods – A classification model is created by training on example human and model generated data

LLMs considered

- ChatGPT (built on GPT-3.5)
- ChatGPT-turbo – latest iteration at the time of the paper
- ChatGLM: based on GLM
- Dolly
- GPT4All
- StableLM

MGT Detection metrics

- Log-Likelihood: measures the token-wise log probability
- Rank: Averaging the rank of each word

$$\text{rank in } p_{\text{det}}(X_i|X_{1:i-1})$$

- Log-Rank: Applies log to rank value of each word
- Entropy: Averages entropy value of each word

$$-\sum_w p_{\text{det}}(X_i = w|X_{1:i-1}) \log p_{\text{det}}(X_i = w|X_{1:i-1})$$

- GLTR: Tool to help annotate whether a text was generated by a model
- DetectGPT: The change of a model's log probability after minor changes are made

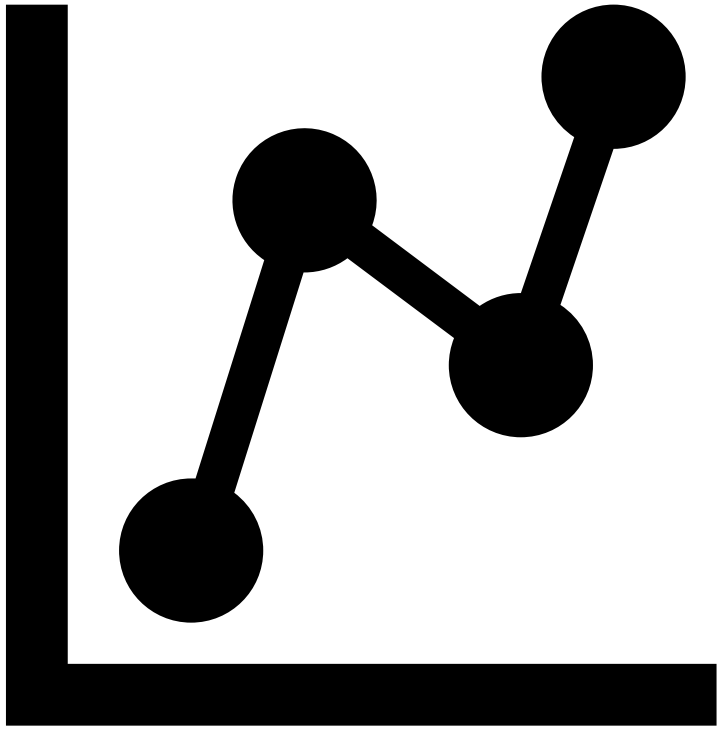
Classification models

- OpenAI Detector: fine-tuning a RoBERTa model using outputs from the GPT2 model with 1.5B parameters
- **ChatGPT Detector**: fine-tuning a RoBERTa model using HC3 data set
- GPTZero: Uses perplexity and burstiness. Modified by authors to make aligned
- **LM Detector**: Fine tuning BERT with an extra classification layer
  - Cane be done with any pre-trained LM

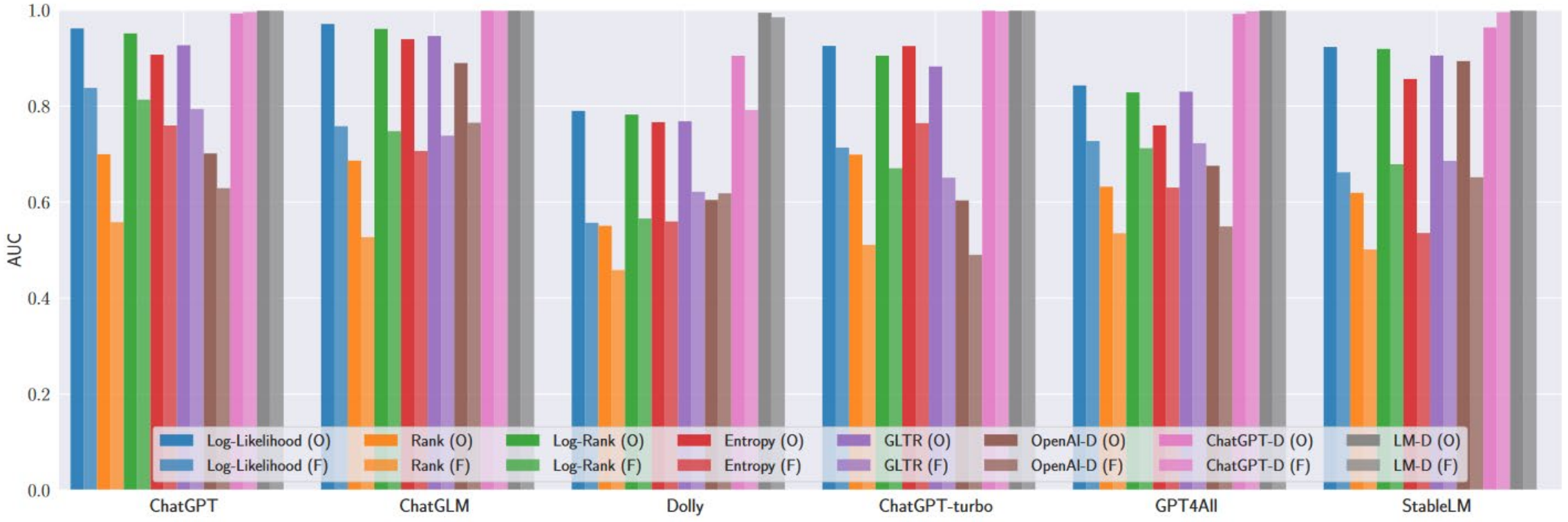
Data sets

- TruthfulQA: 817 questions
- SQuAD1: Over 100,000 questions
  - Randomly sampled 1,000
- NarrativeQA: 1,567 stories and 46,765 questions
  - Randomly sampled 1,000

Evaluation



Evaluation: How to not present data



Evaluation: F1 Score

- $Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ All\ Positives}$
- $Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ Samples\ that\ should\ have\ been\ positive}$

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

- *Can be misleading for class-imbalanced tests*
  - Other evaluation metrics include accuracy, precision, recall, and AUC

Evaluation: MGT Detection

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	Log-Likelihood	0.921	0.945	0.735	0.853	0.769	0.875
	Rank	0.709	0.677	0.646	0.711	0.653	0.668
	Log-Rank	0.903	0.914	0.726	0.850	0.754	0.875
	Entropy	0.824	0.882	0.717	0.855	0.707	0.789
	GLTR	0.882	0.908	0.752	0.848	0.783	0.838
	DetectGPT	0.874	0.899	0.701	0.836	0.713	0.830
	GPTZero-align	0.675	0.839	0.480	0.587	0.368	0.498
	OpenAI-D	0.639	0.697	0.603	0.565	0.653	0.721
	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	Log-Likelihood	0.736	0.688	0.748	0.742	0.771	0.792
	Rank	0.628	0.647	0.647	0.599	0.610	0.711
	Log-Rank	0.707	0.689	0.737	0.732	0.749	0.795
	Entropy	0.725	0.681	0.717	0.689	0.749	0.797
	GLTR	0.728	0.710	0.745	0.734	0.780	0.812
	DetectGPT	0.562	0.400	0.549	0.458	0.604	0.725
	GPTZero-align	0.906	0.866	0.742	0.905	0.775	0.593
	OpenAI-D	0.395	0.500	0.453	0.432	0.473	0.484
	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	Log-Likelihood	0.725	0.646	0.688	0.667	0.812	0.821
	Rank	0.643	0.609	0.640	0.598	0.690	0.719
	Log-Rank	0.724	0.640	0.685	0.661	0.789	0.828
	Entropy	0.713	0.630	0.725	0.673	0.805	0.789
	GLTR	0.706	0.673	0.695	0.651	0.792	0.807
	DetectGPT	0.519	0.455	0.526	0.465	0.684	0.705
	GPTZero-align	0.707	0.722	0.464	0.798	0.382	0.563
	OpenAI-D	0.384	0.414	0.527	0.369	0.438	0.566
	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: MGT Detection

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	Log-Likelihood	0.921	0.945	0.735	0.853	0.769	0.875
	Rank	0.709	0.677	0.646	0.711	0.653	0.668
	Log-Rank	0.903	0.914	0.726	0.850	0.754	0.875
	Entropy	0.824	0.882	0.717	0.855	0.707	0.789
	GLTR	0.882	0.908	0.752	0.848	0.783	0.838
	DetectGPT	0.874	0.899	0.701	0.836	0.713	0.830
	GPTZero-align	0.675	0.839	0.480	0.587	0.368	0.498
	OpenAI-D	0.639	0.697	0.603	0.565	0.653	0.721
	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	Log-Likelihood	0.736	0.688	0.748	0.742	0.771	0.792
	Rank	0.628	0.647	0.647	0.599	0.610	0.711
	Log-Rank	0.707	0.689	0.737	0.732	0.749	0.795
	Entropy	0.725	0.681	0.717	0.689	0.749	0.797
	GLTR	0.728	0.710	0.745	0.734	0.780	0.812
	DetectGPT	0.562	0.400	0.549	0.458	0.604	0.725
	GPTZero-align	0.906	0.866	0.742	0.905	0.775	0.593
	OpenAI-D	0.395	0.500	0.453	0.432	0.473	0.484
	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	Log-Likelihood	0.725	0.646	0.688	0.667	0.812	0.821
	Rank	0.643	0.609	0.640	0.598	0.690	0.719
	Log-Rank	0.724	0.640	0.685	0.661	0.789	0.828
	Entropy	0.713	0.630	0.725	0.673	0.805	0.789
	GLTR	0.706	0.673	0.695	0.651	0.792	0.807
	DetectGPT	0.519	0.455	0.526	0.465	0.684	0.705
	GPTZero-align	0.707	0.722	0.464	0.798	0.382	0.563
	OpenAI-D	0.384	0.414	0.527	0.369	0.438	0.566
	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: MGT Detection

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	Log-Likelihood	0.921	0.945	0.735	0.853	0.769	0.875
	Rank	0.709	0.677	0.646	0.711	0.653	0.668
	Log-Rank	0.903	0.914	0.726	0.850	0.754	0.875
	Entropy	0.824	0.882	0.717	0.855	0.707	0.789
	GLTR	0.882	0.908	0.752	0.848	0.783	0.838
	DetectGPT	0.874	0.899	0.701	0.836	0.713	0.830
	GPTZero-align	0.675	0.839	0.480	0.587	0.368	0.498
	OpenAI-D	0.639	0.697	0.603	0.565	0.653	0.721
	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	Log-Likelihood	0.736	0.688	0.748	0.742	0.771	0.792
	Rank	0.628	0.647	0.647	0.599	0.610	0.711
	Log-Rank	0.707	0.689	0.737	0.732	0.749	0.795
	Entropy	0.725	0.681	0.717	0.689	0.749	0.797
	GLTR	0.728	0.710	0.745	0.734	0.780	0.812
	DetectGPT	0.562	0.400	0.549	0.458	0.604	0.725
	GPTZero-align	0.906	0.866	0.742	0.905	0.775	0.593
	OpenAI-D	0.395	0.500	0.453	0.432	0.473	0.484
	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	Log-Likelihood	0.725	0.646	0.688	0.667	0.812	0.821
	Rank	0.643	0.609	0.640	0.598	0.690	0.719
	Log-Rank	0.724	0.640	0.685	0.661	0.789	0.828
	Entropy	0.713	0.630	0.725	0.673	0.805	0.789
	GLTR	0.706	0.673	0.695	0.651	0.792	0.807
	DetectGPT	0.519	0.455	0.526	0.465	0.684	0.705
	GPTZero-align	0.707	0.722	0.464	0.798	0.382	0.563
	OpenAI-D	0.384	0.414	0.527	0.369	0.438	0.566
	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: MGT Detection

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: MGT Detection

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: MGT Detection

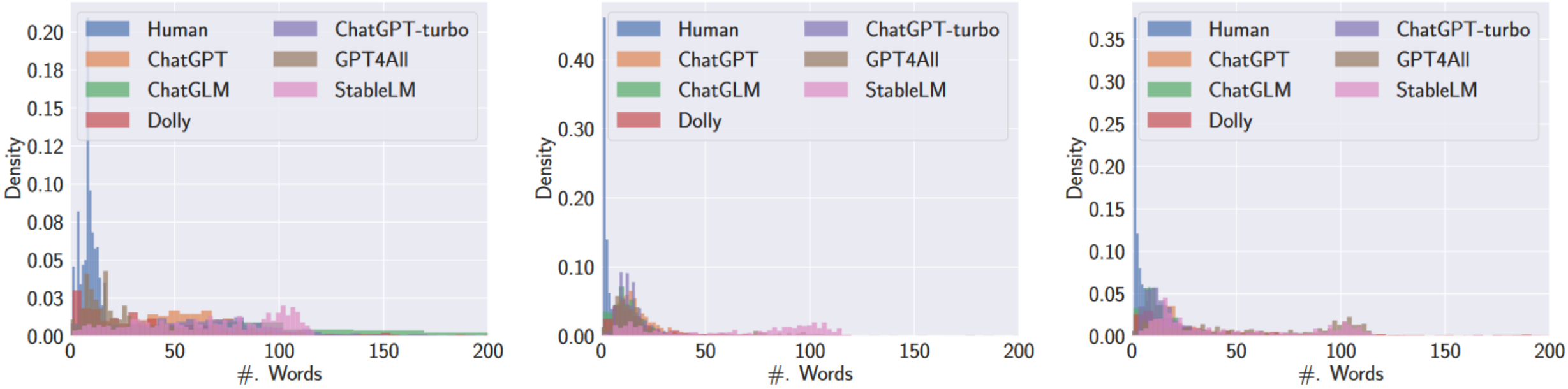
Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	ChatGPT-D	0.974	0.987	0.746	0.967	0.946	0.778
	LM-D	**1.000**	**0.997**	**0.966**	**0.997**	**1.000**	**1.000**
SQuAD1	ChatGPT-D	0.843	0.777	0.679	0.863	0.821	0.784
	LM-D	**0.989**	**0.960**	**0.939**	**0.985**	**0.981**	**0.996**
NarrativeQA	ChatGPT-D	0.813	0.804	0.669	0.787	0.748	0.784
	LM-D	**0.948**	**0.931**	**0.875**	**0.946**	**0.987**	**0.977**

Evaluation: Time Cost

Table 3: Time cost (seconds) to differentiate texts generated by ChatGPT or humans.

Method	TruthfulQA	SQuAD1	NarrativeQA
Log-Likelihood	11	9	10
Rank	12	9	10
Log-Rank	12	8	10
Entropy	11	8	9
GLTR	12	9	10
GPTZero	275	347	350
DetectGPT	877	629	577
OpenAI-D	**5**	3	4
ChatGPT-D	**5**	**2**	**3**
LM-D	25	10	16

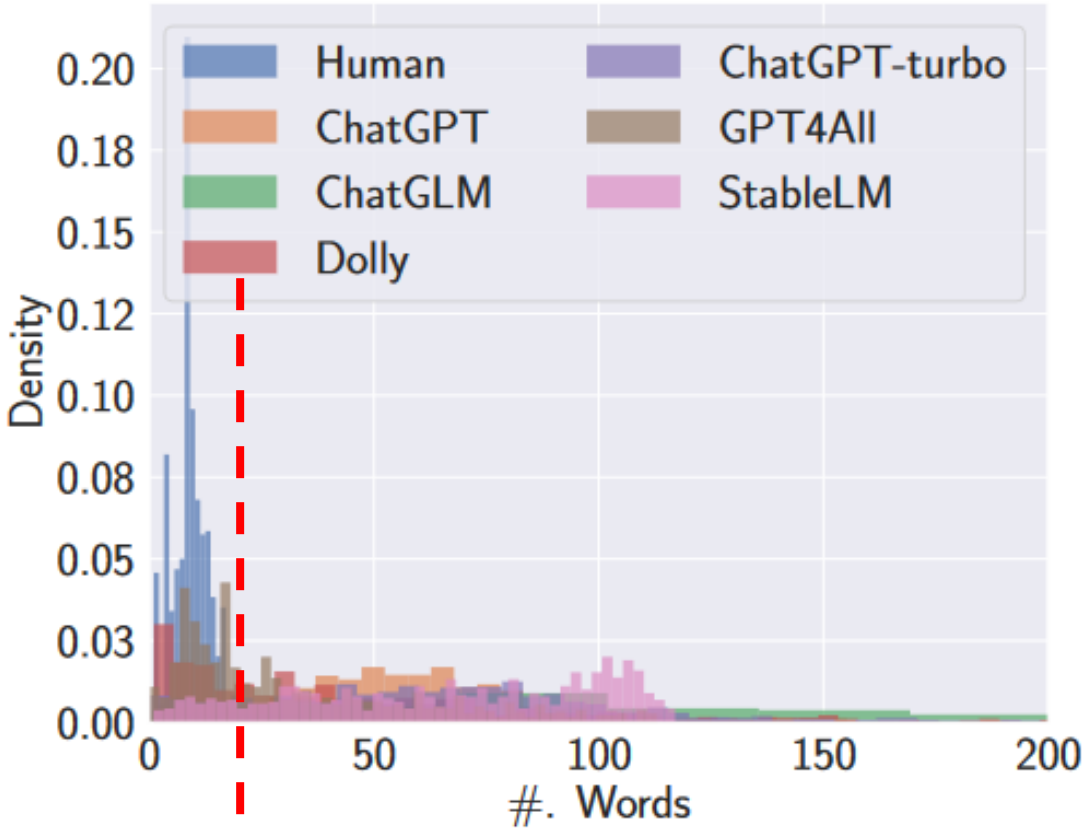
Ablation: Number of Words



(a) TruthfulQA (b) SQuAD1 (c) NarrativeQA

Figure 1: The distribution of #. words in human-written texts and machine-generated texts.

Ablation: Number of Words



(a) TruthfulQA

Ablation: Number of Words

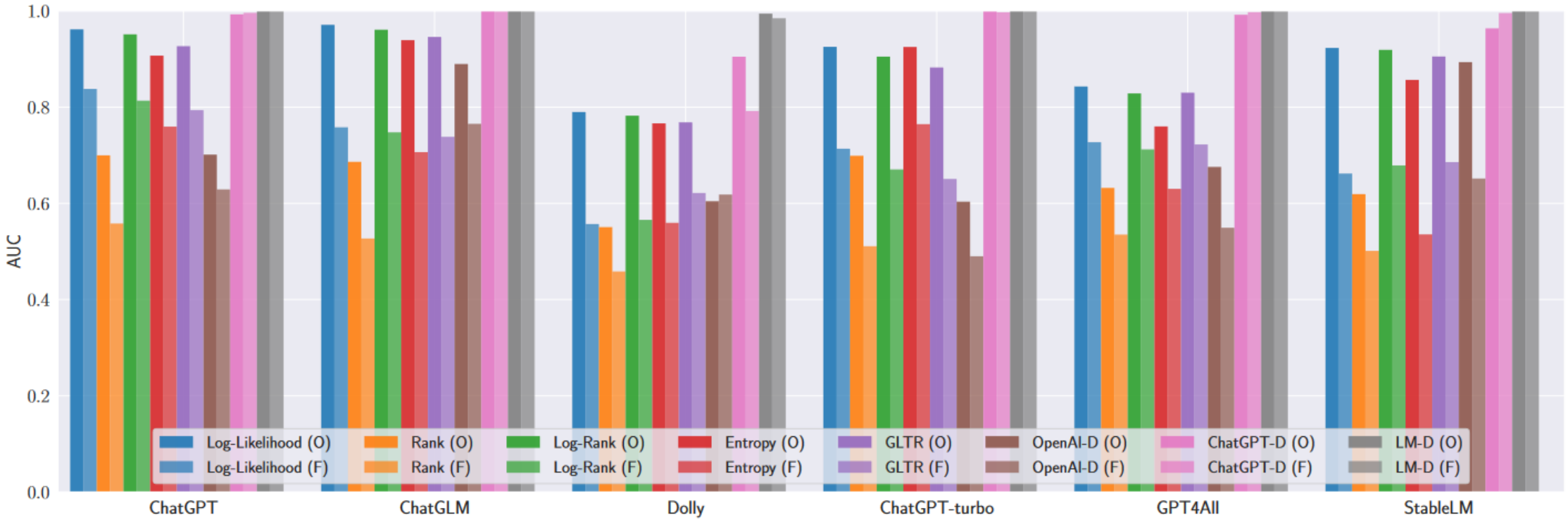
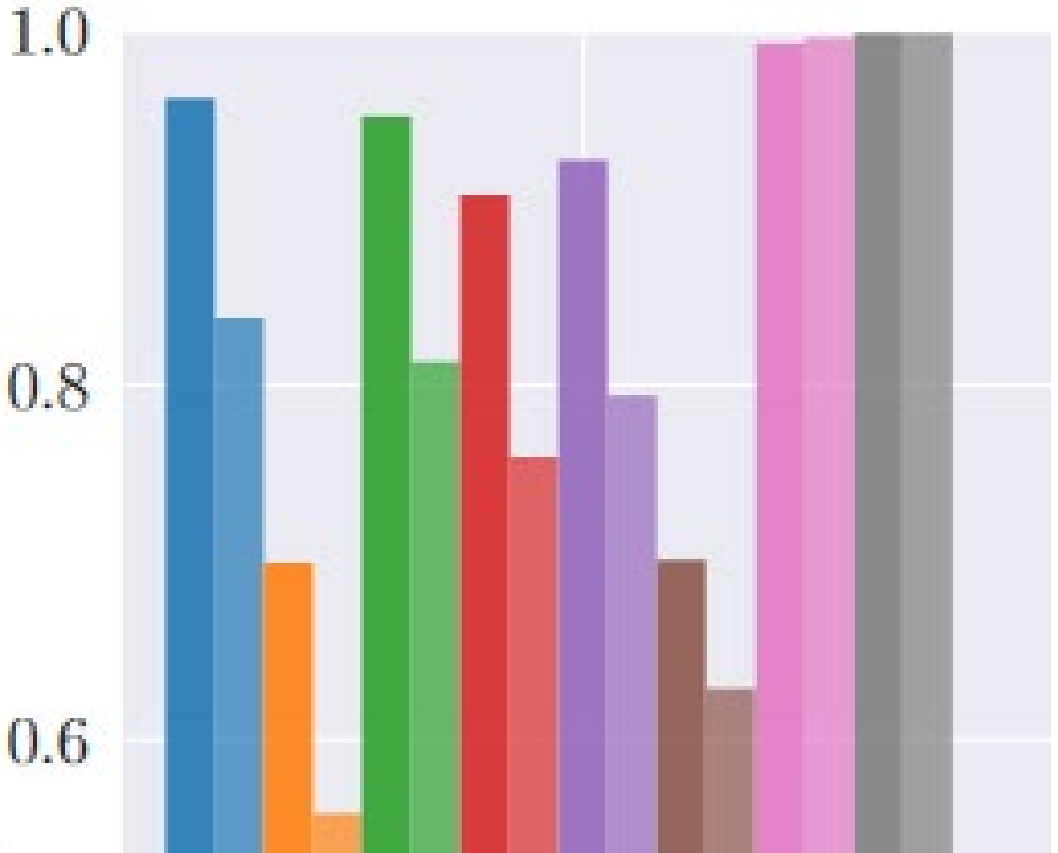
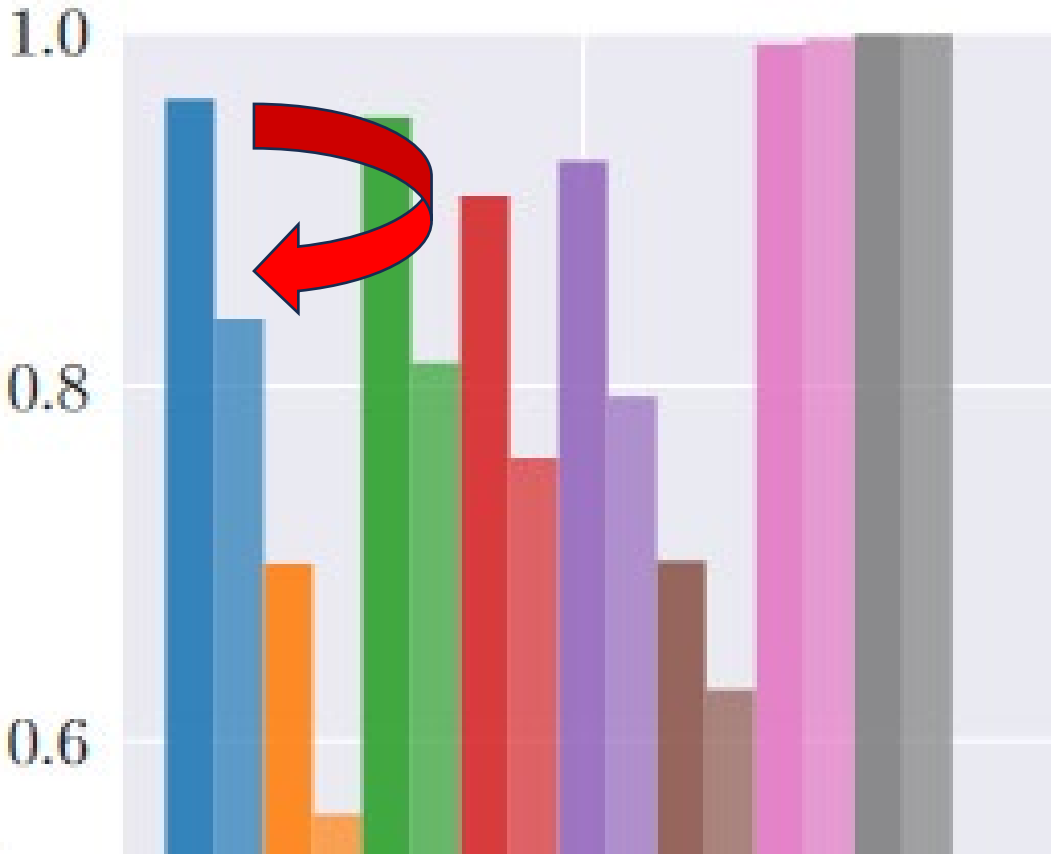


Figure 2: The AUC of different detection methods against different LLMs on TruthfulQA. Note that, for each detection method, (O) denotes the performance on the original dataset and (F) denotes the performance on the filtered dataset that only contains texts with no more than 25 words.

Ablation: Number of Words



Ablation: Number of Words



LLM as a Detector

- Prompts: Judge whether the sentence is generated by human or machine: <sentence>, and please only answer “human” or “machine”

ChatGPT-turbo as a Detector

- Prompts: Judge whether the sentence is generated by human or machine: <sentence>, and please only answer “human” or “machine”

ChatGPT	ChatGLM	Dolly
0.406	0.325	0.338
ChatGPT-turbo	**GPT4All**	**StableLM**
0.524	0.360	0.409

ChatGPT-turbo as a Detector

- Prompts: Judge whether the sentence is generated by human or machine: <sentence>, and please only answer “human” or “machine”

ChatGPT	ChatGLM	Dolly
0.406	0.325	0.338
ChatGPT-turbo	**GPT4All**	**StableLM**
0.524	0.360	0.409

Fine-tune with Fewer Samples

- 10 samples for fine tuning tended to be sufficient

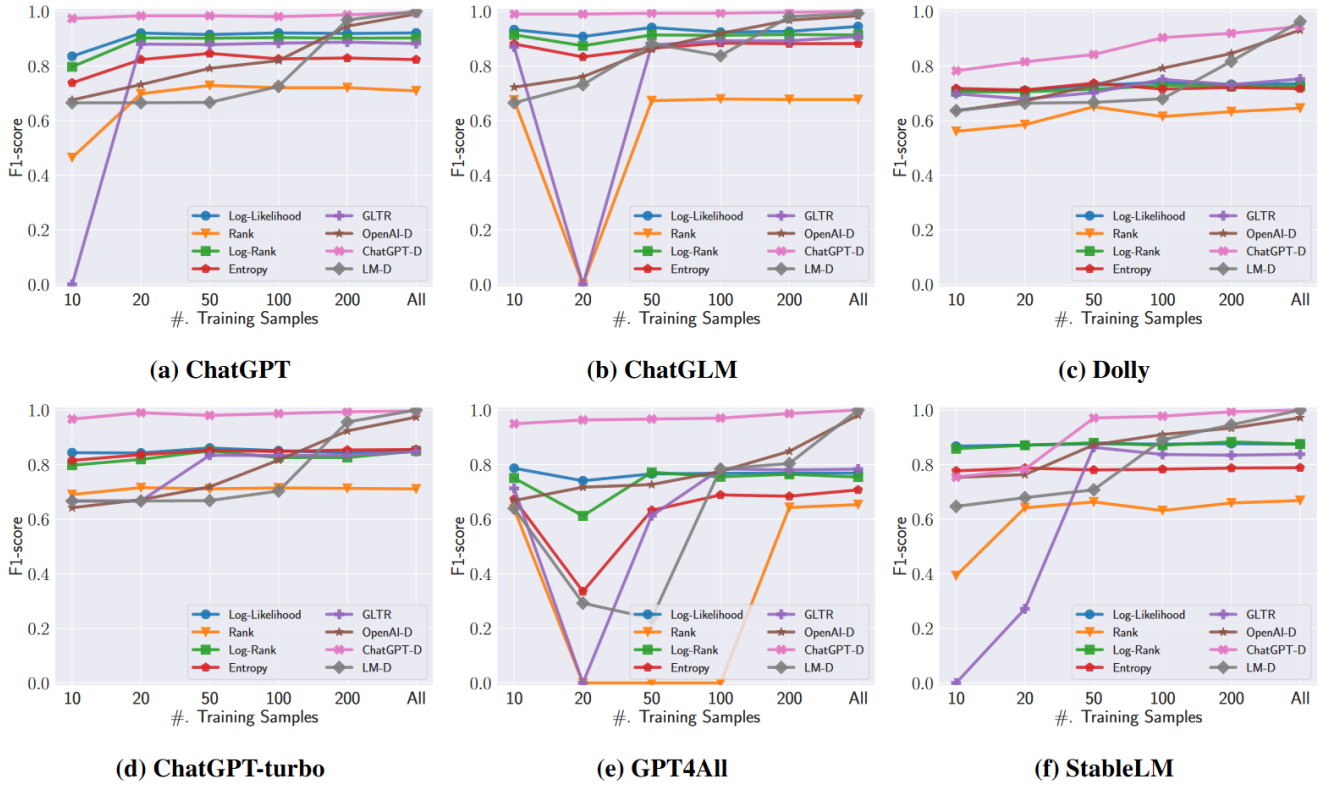
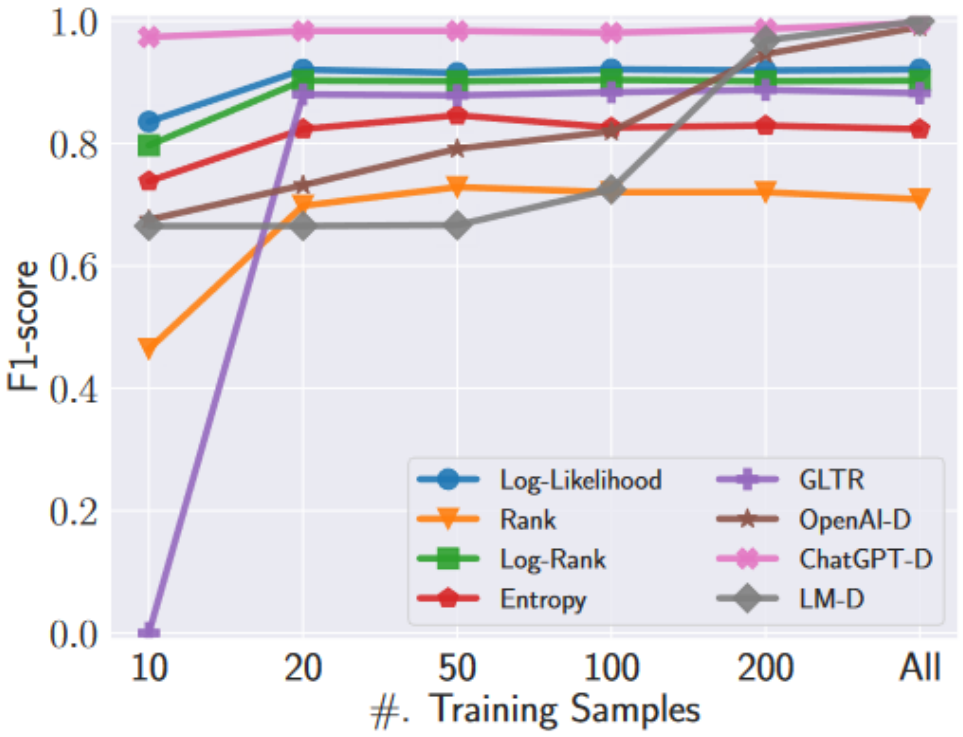


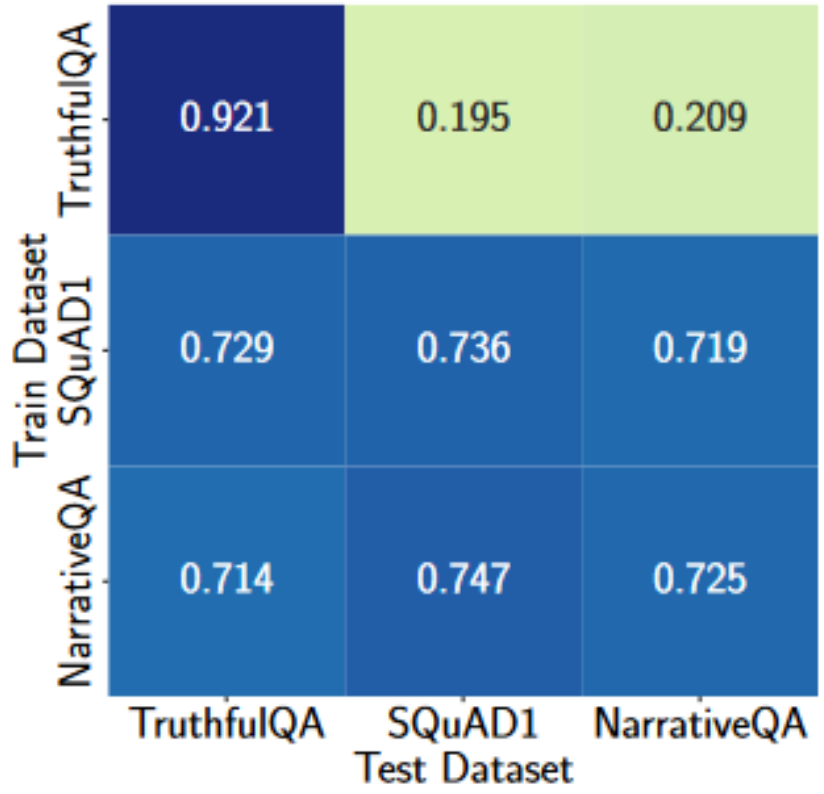
Figure 3: The F1-score of different detection methods with different numbers of training samples on TruthfulQA.

Fine-tune with Fewer Samples



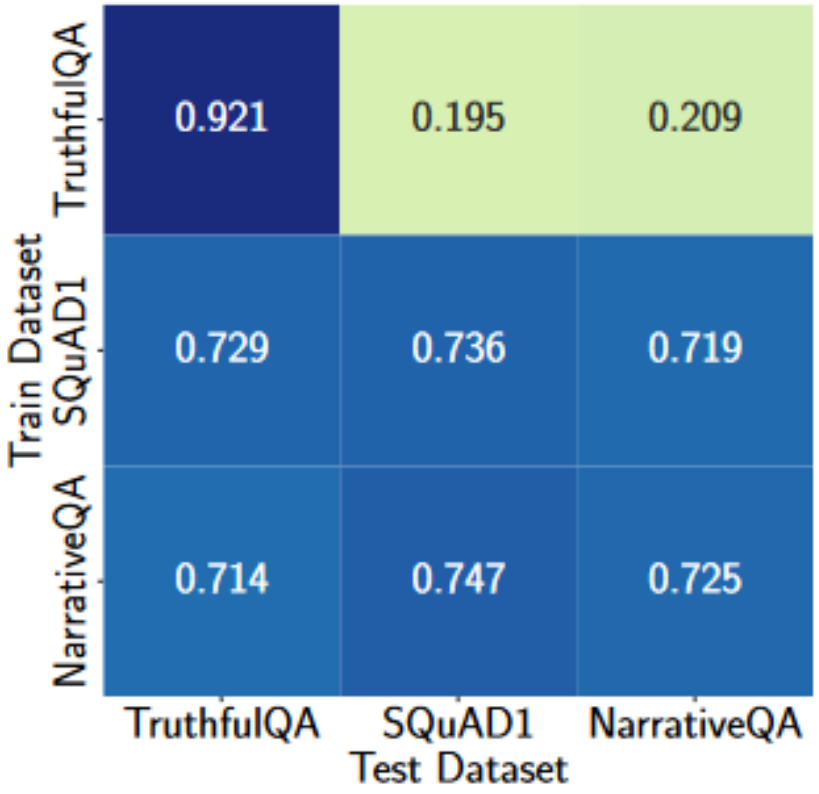
(a) ChatGPT

Transfer Setting Using Different Datasets For Training



(a) Log-Likelihood

Transfer Setting Using Different Datasets For Training



(a) Log-Likelihood

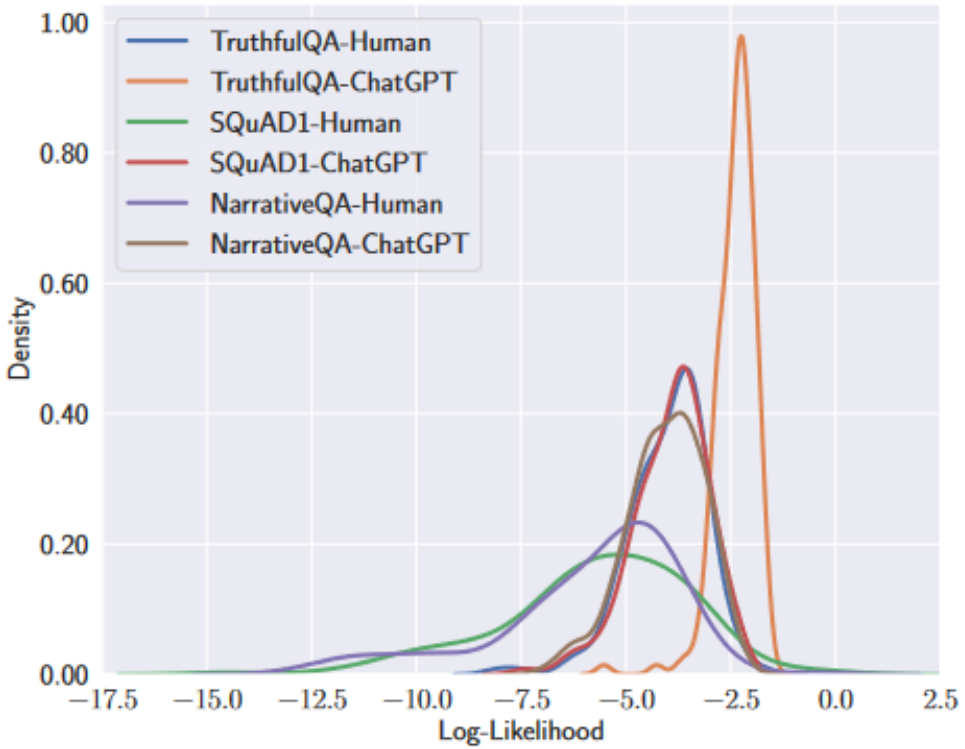
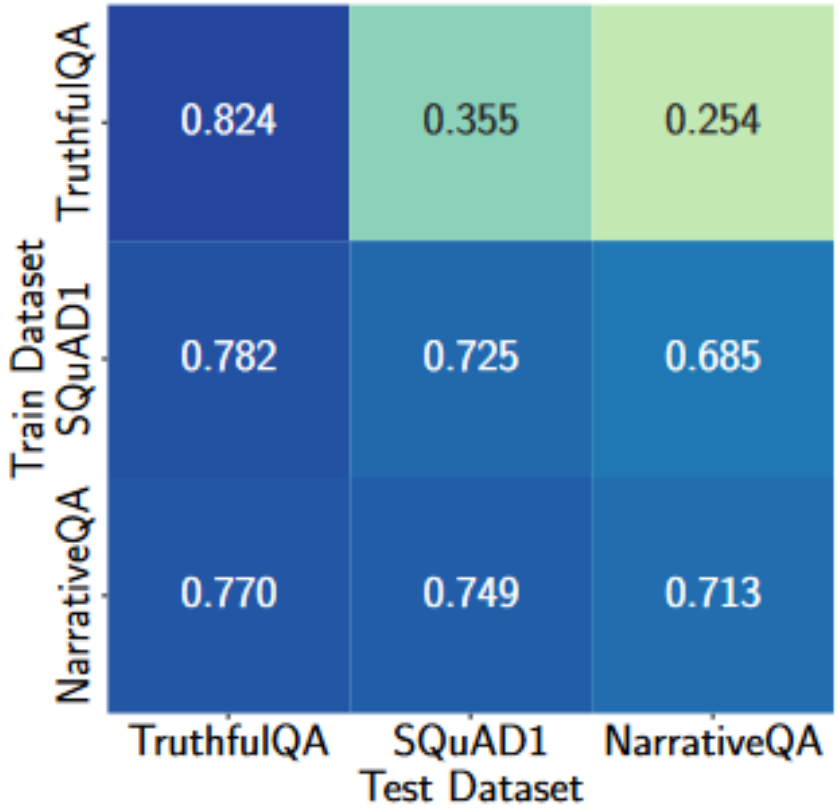
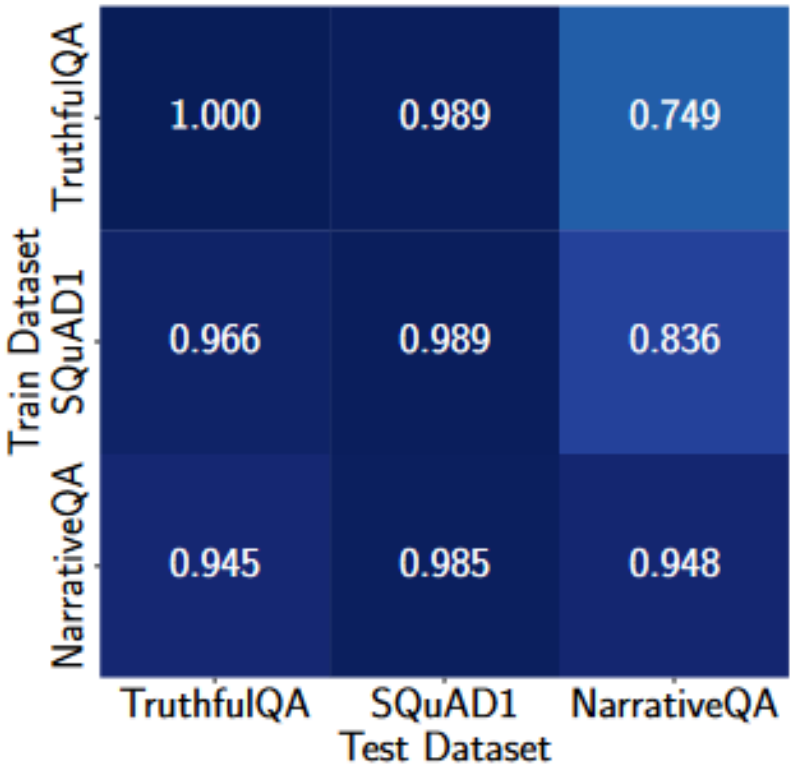


Figure 5: The F1-score of different detection methods on the text attribution task.

Transfer Setting Using Different Datasets For Training: Metric vs Model

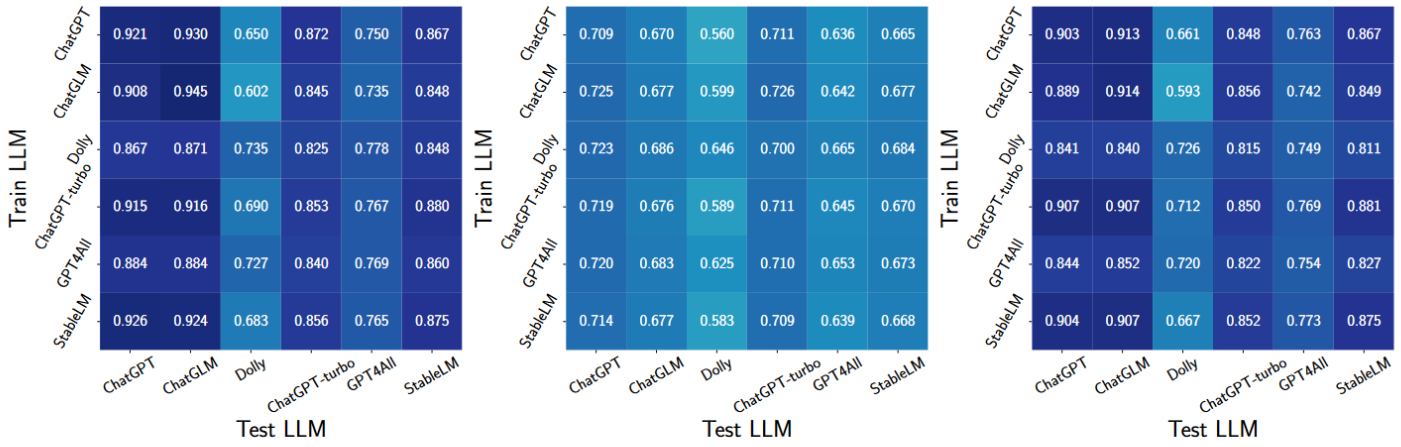


(d) Entropy

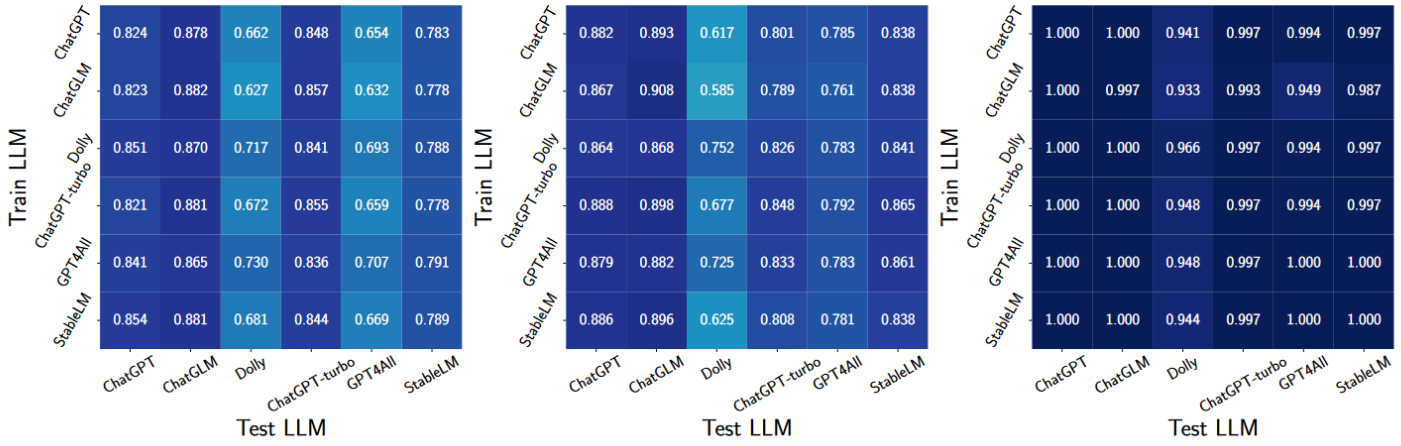


(f) LM-D

Transfer Setting: Training to Detect One LLM and Applying it to Another



(a) Log-Likelihood (b) Rank (c) Log-Rank



(d) Entropy (e) GLTR (f) LM-D

Figure 6: The F1-score of different detection methods on TruthfulQA when the train LLM and the test LLM are different.

Text Attribution by LLM

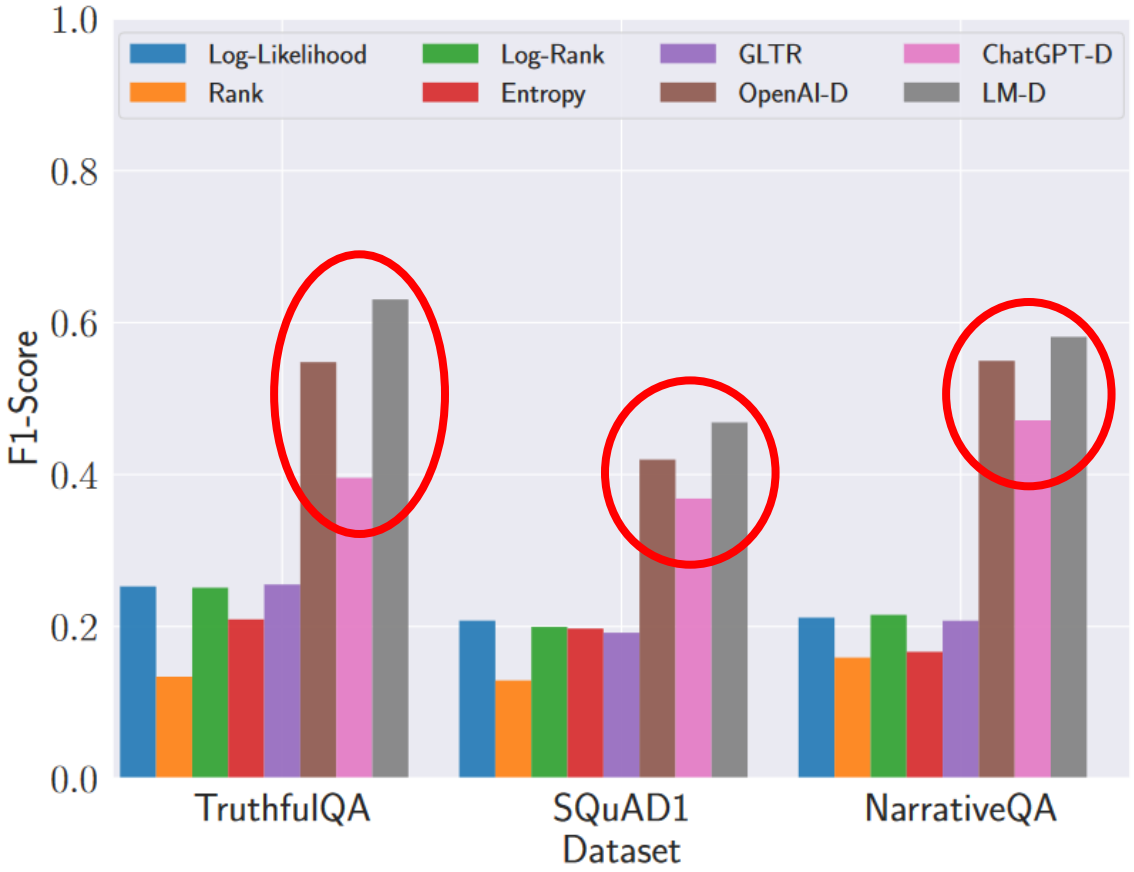
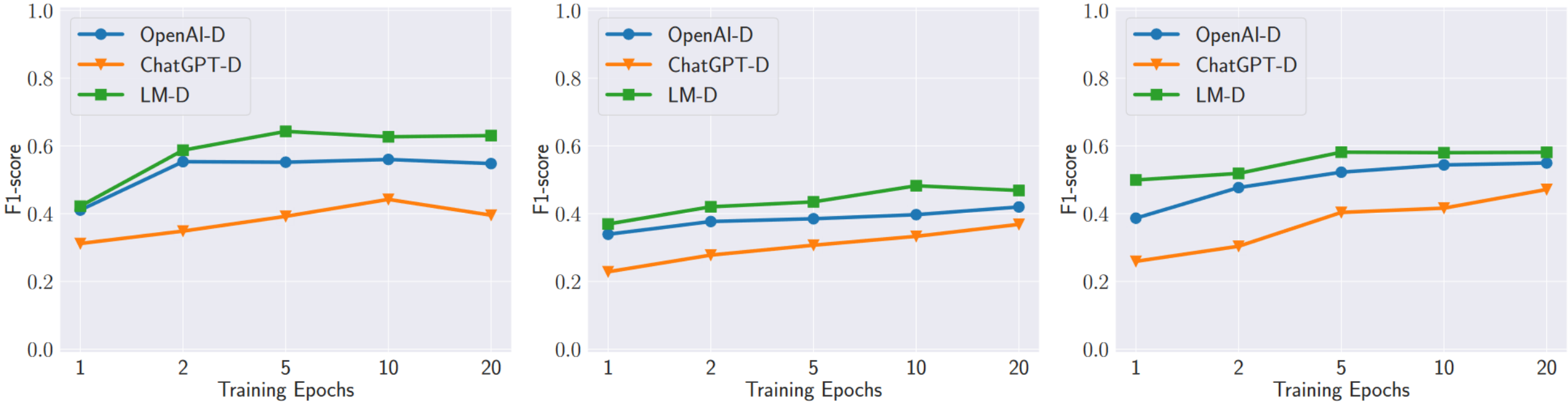


Figure 7: The F1-score of different detection methods on the text attribution task.

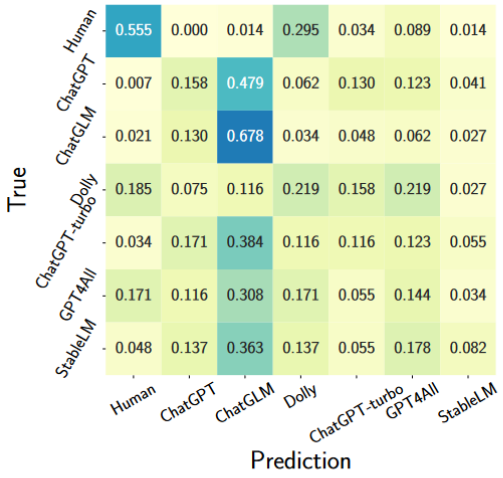
Text Attribution by LLM: Epochs?



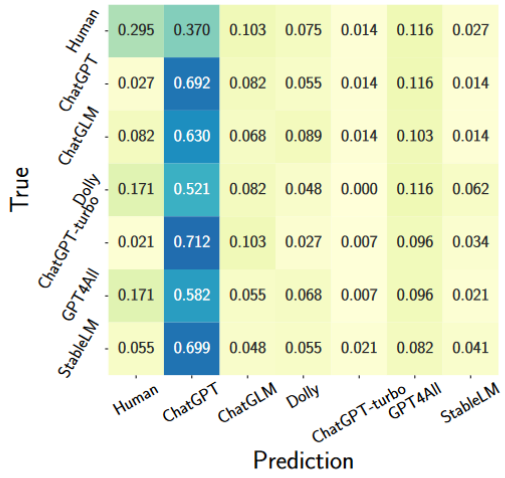
(a) TruthfulQA **(b) SQuAD1** **(c) NarrativeQA**

Figure 8: The F1-score of text attribution performance with different training epochs.

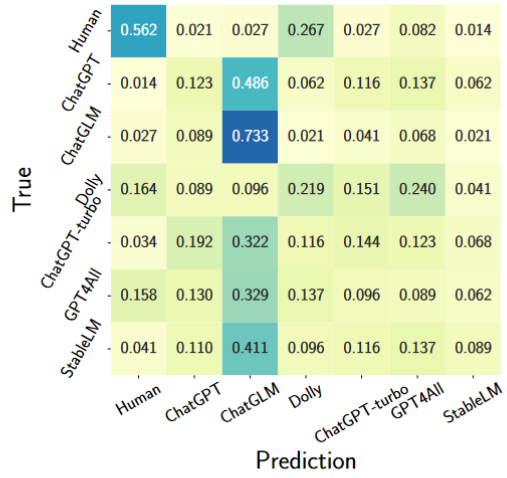
Text Attribution by LLM:LLM Breakdown



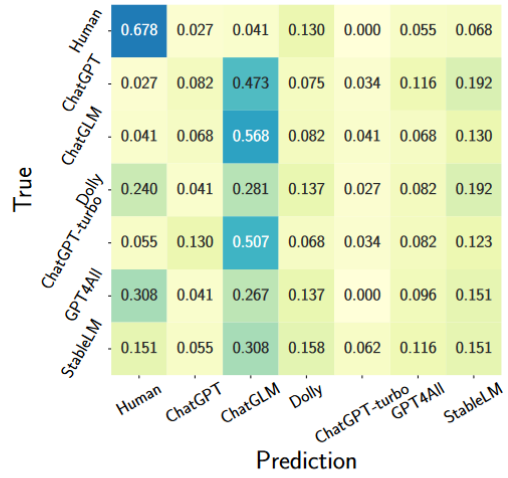
(a) Log-Likelihood



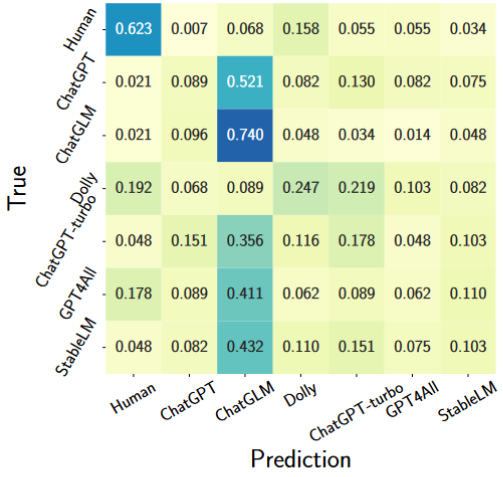
(b) Rank



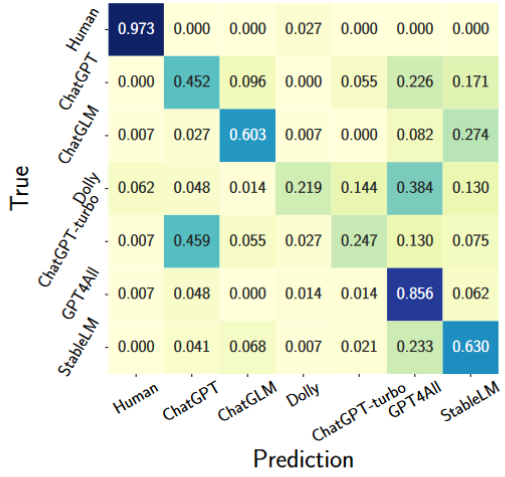
(c) Log-Rank



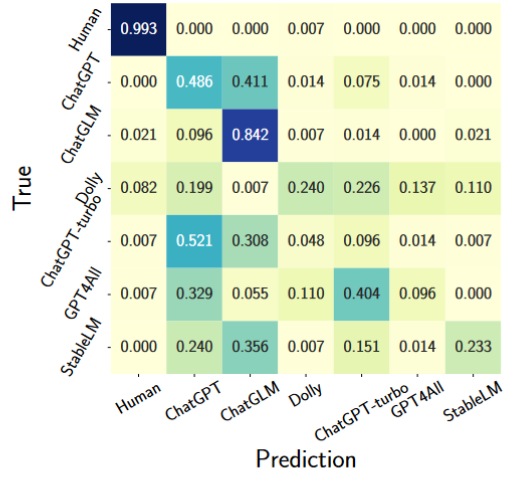
(d) Entropy



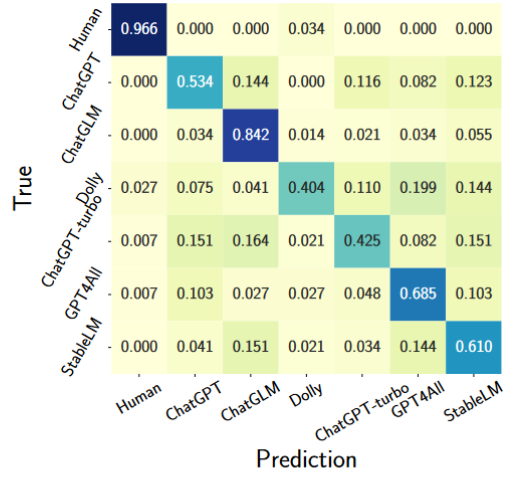
(e) GLTR



(f) OpenAI-D

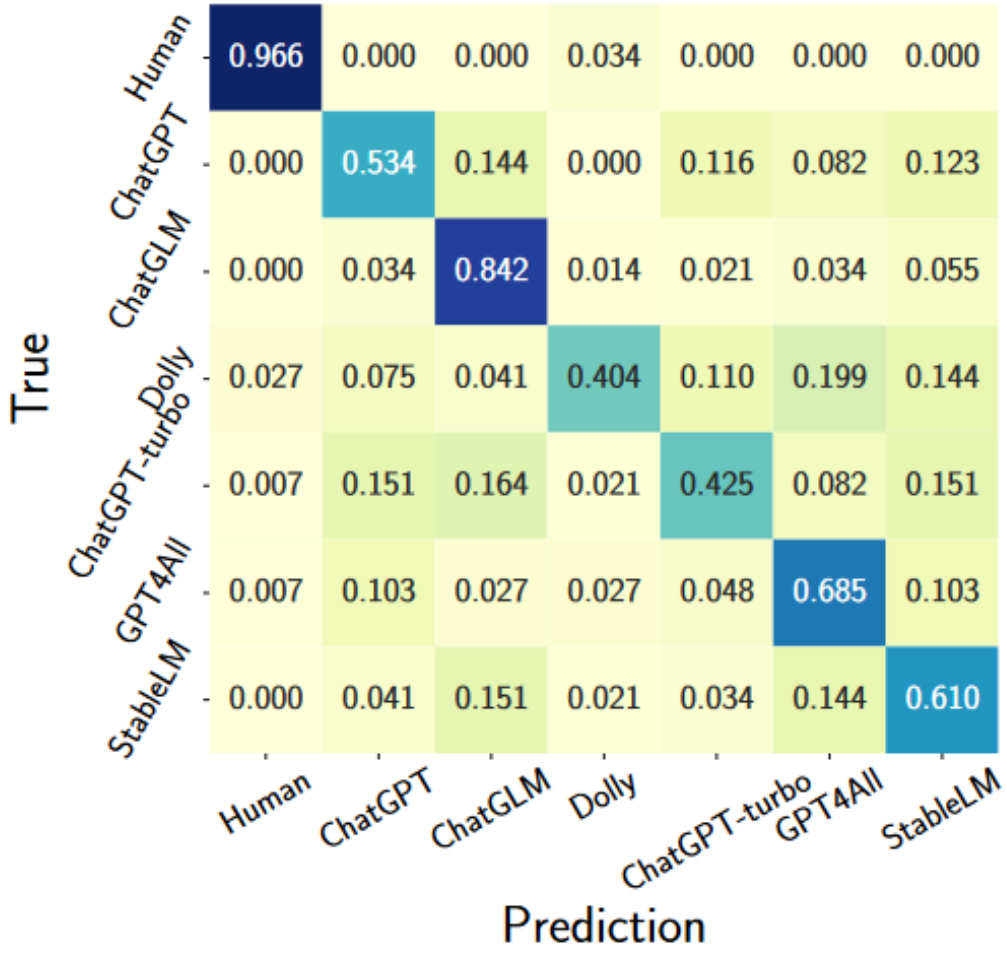


(g) ChatGPT-D



(h) LM-D

Text Attribution by LLM:LLM Breakdown



(h) LM-D

Adversarial Attacks

- Select 100 MGTs that are correctly classified by the detection method
- Use TextAttack to create an adversarial attack

Table 5: The attack accuracy on the texts being classified as MGTs.

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	ChatGPT-D	0.990	1.000	1.000	1.000	0.990	0.980
	LM-D	0.000	0.000	0.080	0.000	0.100	0.020
SQuAD1	ChatGPT-D	0.990	0.990	0.980	1.000	0.990	1.000
	LM-D	0.030	0.070	0.020	0.060	0.030	0.020
NarrativeQA	ChatGPT-D	1.000	1.000	0.980	1.000	0.990	1.000
	LM-D	0.430	0.540	0.440	0.550	0.330	0.450

- Trained on the “same distribution dataset”

Adversarial Attacks

- Select 100 MGTs that are correctly classified by the detection method
- Use TextAttack to create an adversarial attack

Table 5: The attack accuracy on the texts being classified as MGTs.

Dataset	Method	ChatGPT	ChatGLM	Dolly	ChatGPT-turbo	GPT4All	StableLM
TruthfulQA	ChatGPT-D	0.990	1.000	1.000	1.000	0.990	0.980
	LM-D	0.000	0.000	0.080	0.000	0.100	0.020
SQuAD1	ChatGPT-D	0.990	0.990	0.980	1.000	0.990	1.000
	LM-D	0.030	0.070	0.020	0.060	0.030	0.020
NarrativeQA	ChatGPT-D	1.000	1.000	0.980	1.000	0.990	1.000
	LM-D	0.430	0.540	0.440	0.550	0.330	0.450

Takeaway

- The authors created a tool in which different datasets, detectors, and metrics can be used
  - Researchers can compare their own detectors against existing work
  - Novel enough to accepted to a conference?
  - Meaningful Evaluations?
- Point to future areas of focus
  - Why are results the way they are?
  - Text Attribution
  - Robustness Against Adaptive Attacks

Table of Contents