IJCAI JEJU 2024

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
中国科学院大学
University of Chinese Academy of Sciences

Ten Words Only Still Help: Improving Black-Box AI-Generated Text Detection via Proxy-Guided Efficient Re-Sampling

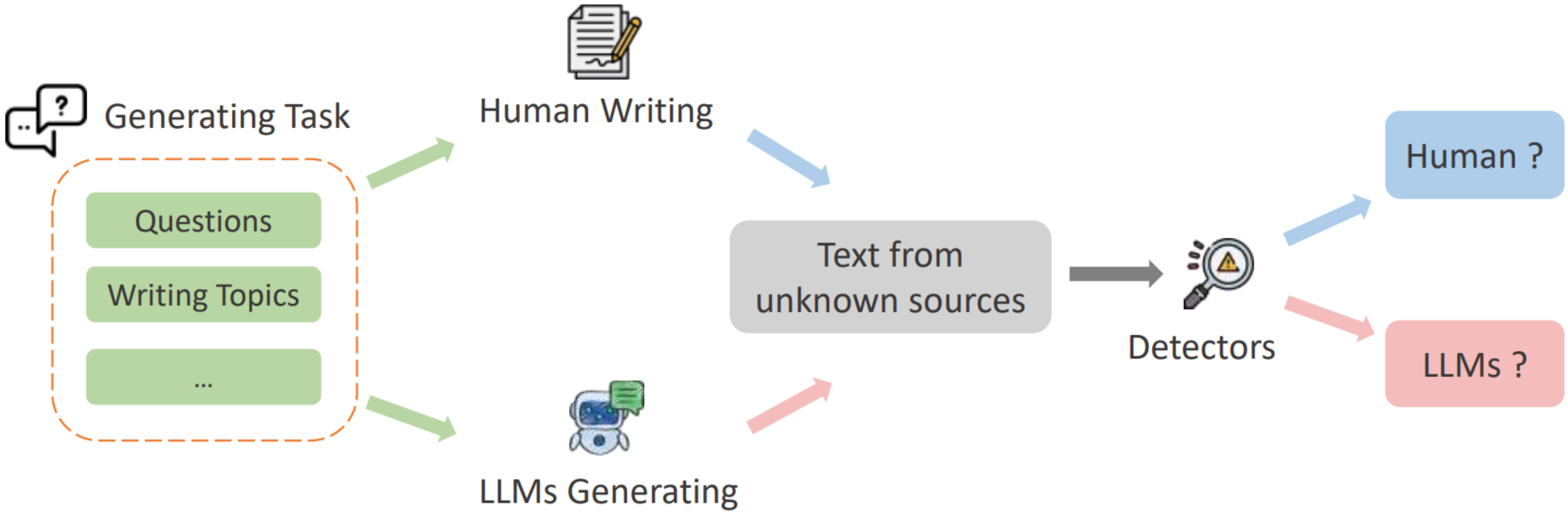
Yuhui Shi[1,2], Qiang Sheng[1], Juan Cao[1,2], Hao Mi[1,2], Beizhe Hu[1,2], Danding Wang[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

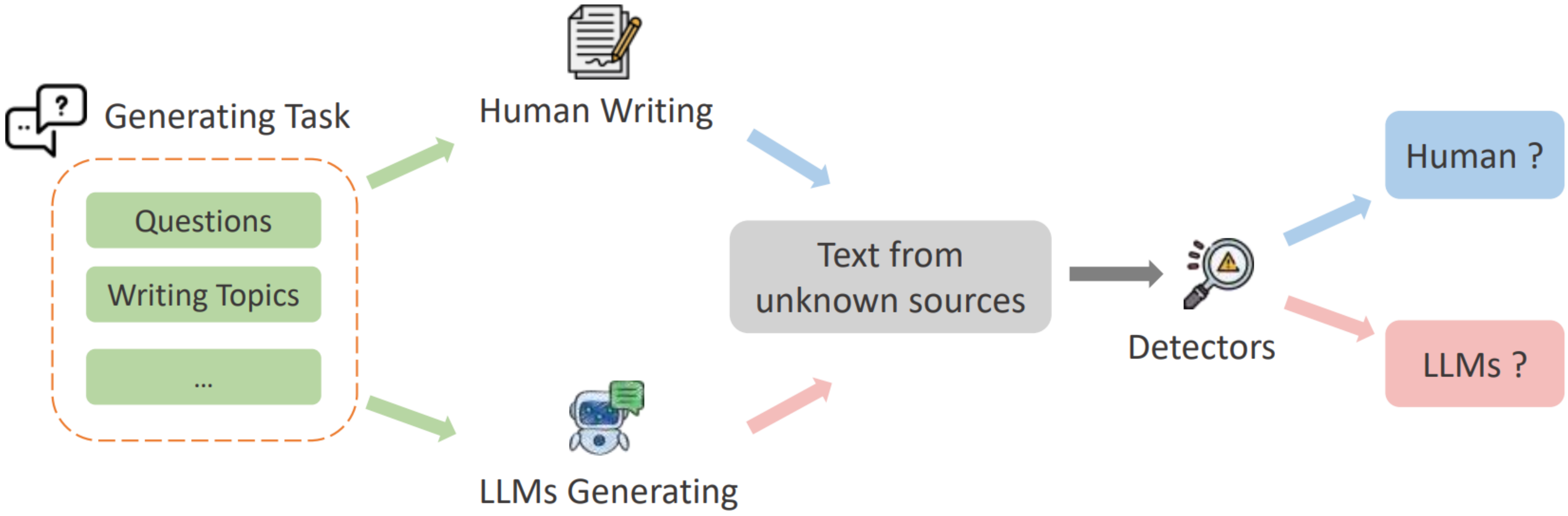
2024.08.09

Introduction



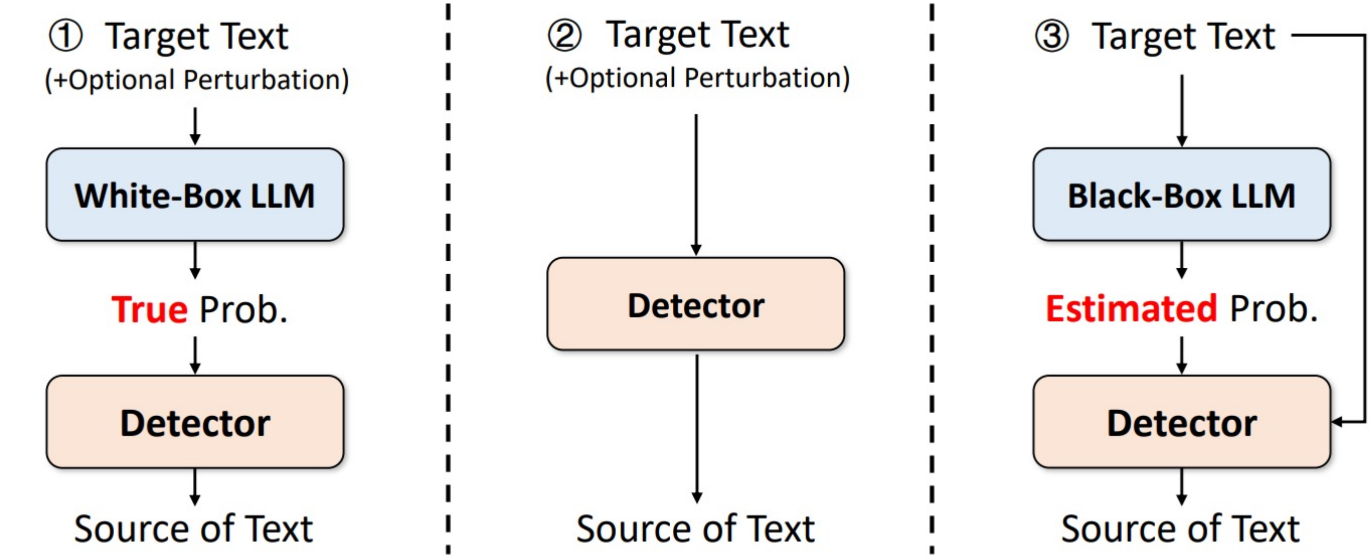
- **Background:** The misuse of large language models (LLMs) has led to issues such as misinformation and academic dishonesty, which makes AI-generated text (AIGT) detection critical.

Introduction



- **Task Formulation:** AIGT detection aims to obtain a classifier $f: \boldsymbol{x} \rightarrow y$, where y is the source of the given text $\boldsymbol{x}$.
  - **Binary AIGT Detection:** $y \in \{\text{human}, \text{AI}\}$
  - **Multiclass AIGT Detection:** $y \in \{\text{human}, \theta_1, \theta_2, \cdots, \theta_M\}$, where θ_i is a LLM

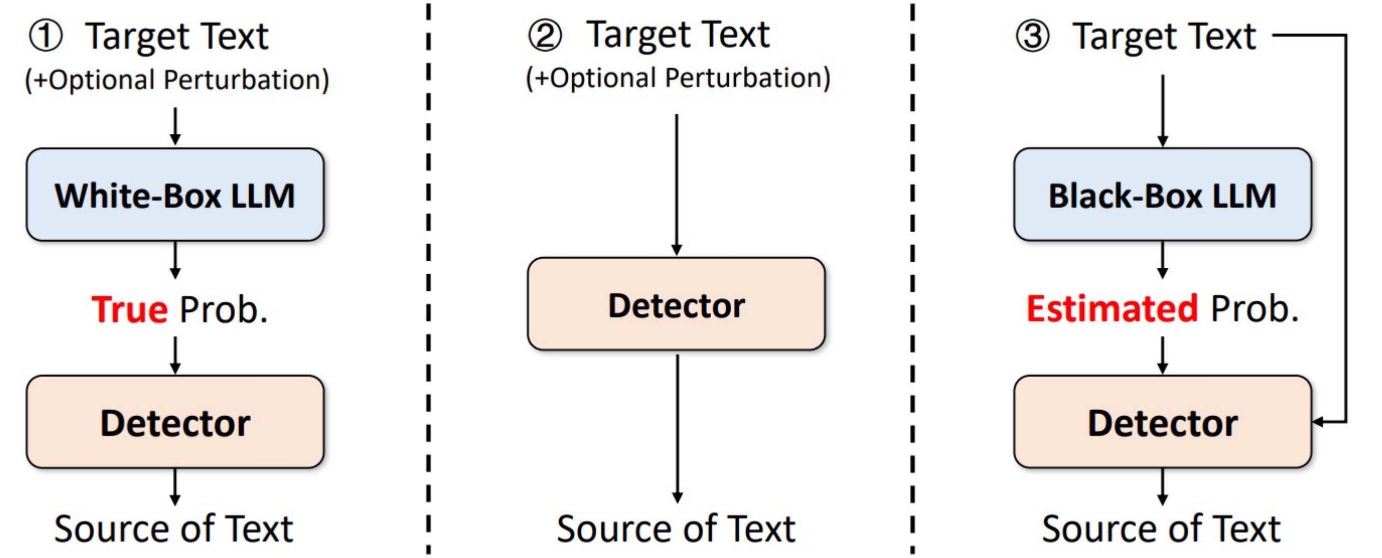
Motivation



	① White-Box Baseline	② Black-Box Baseline	③ POGER (Ours)
LLM Logits Free	✗	✓	✓
Black-Box Performance	✗	–	✓
OOD Generalization	✓	✗	✓

- **Challenge:** White-box methods have better performance and generalizability, but they require access to LLMs' internal states and are not applicable to black-box settings.

Motivation



	① White-Box Baseline	② Black-Box Baseline	③ POGER (Ours)
LLM Logits Free	✗	✓	✓
Black-Box Performance	✗	—	✓
OOD Generalization	✓	✗	✓

➢ **Solution:** Estimate word generation probabilities as pseudo white-box features via multiple re-sampling to help improve AIGT detection under the black-box setting.

Preliminary Study

A naive solution:

a. For each word in given text $\boldsymbol{x}$, we instruct the black-box LLM for N times using the following prompt:

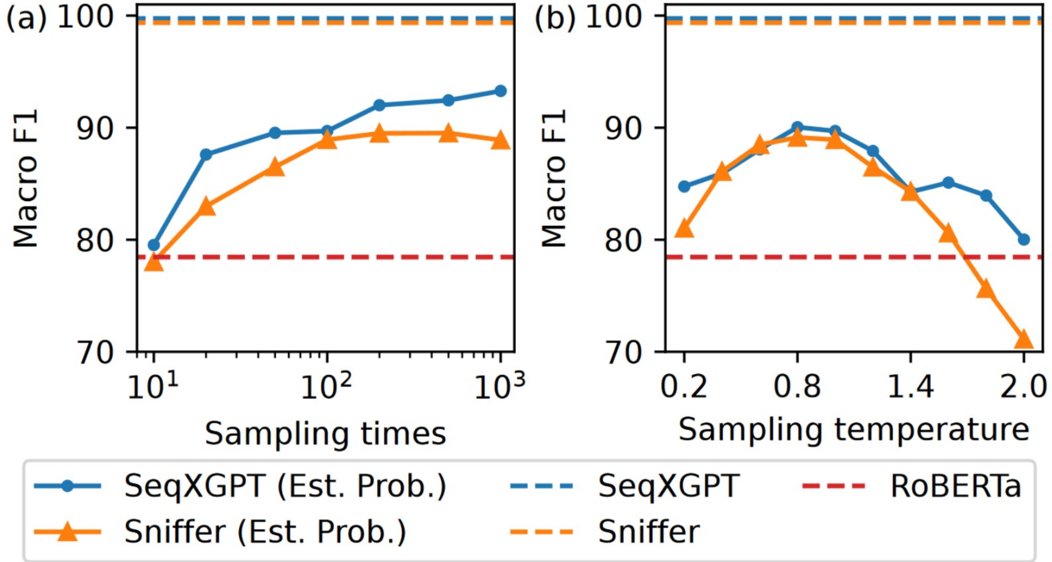
Please continue writing the following text, starting from the next word: $\{x_{<i}\}$.

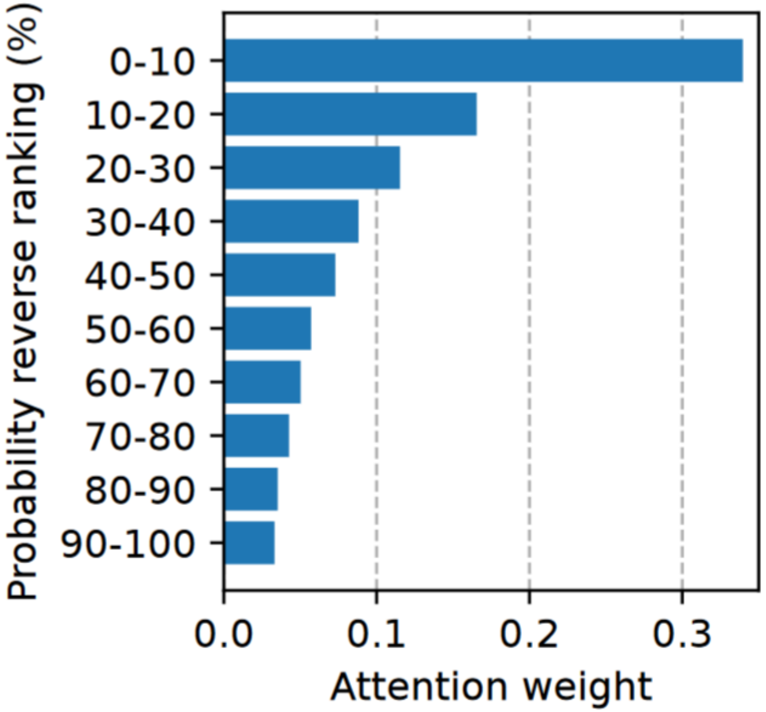
b. The estimated probability of x_i given $\{x_{<i}\}$ is computed as the frequency of x_i in the output word set $\{o_j\}_{j=1}^{N}$:

$$\hat{p}(x_i|x_{<i}) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(o_j = x_i)\,.$$

c. Use estimated probability as an alternative input of white-box methods.

Preliminary Study



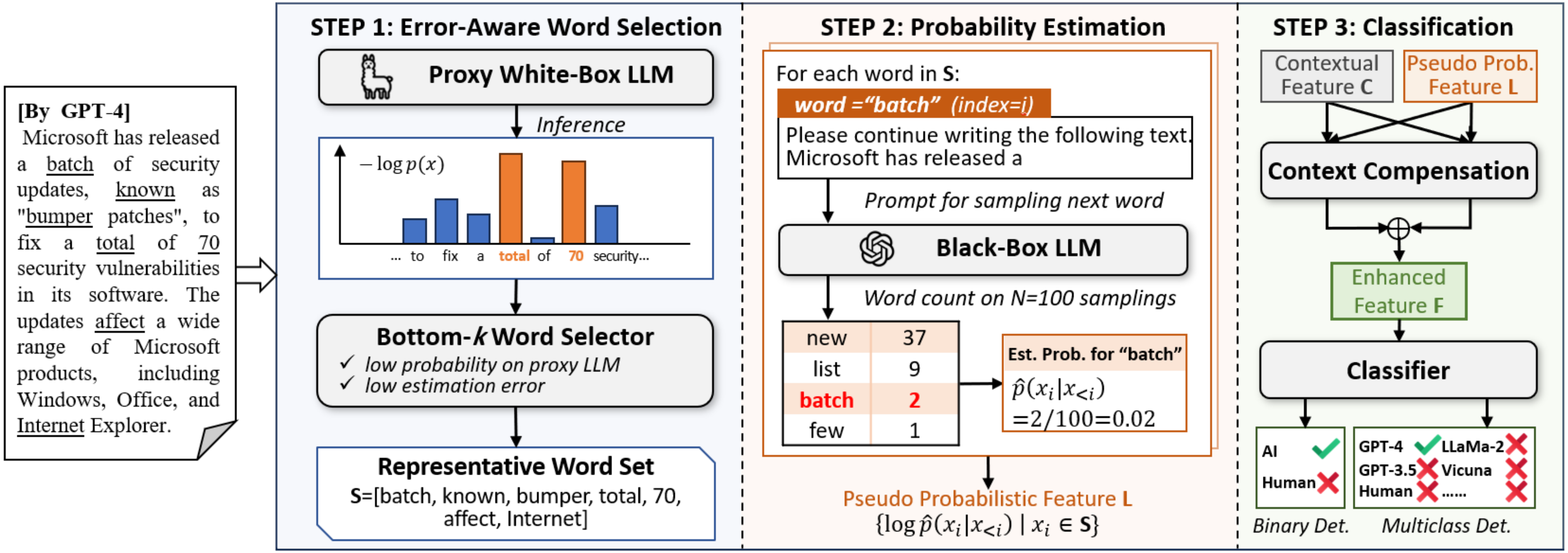


Finding 1: It is feasible to perform black-box AIGT detection by estimated probs.

Finding 2: Low-probability words gain higher attention from the detector.

Method

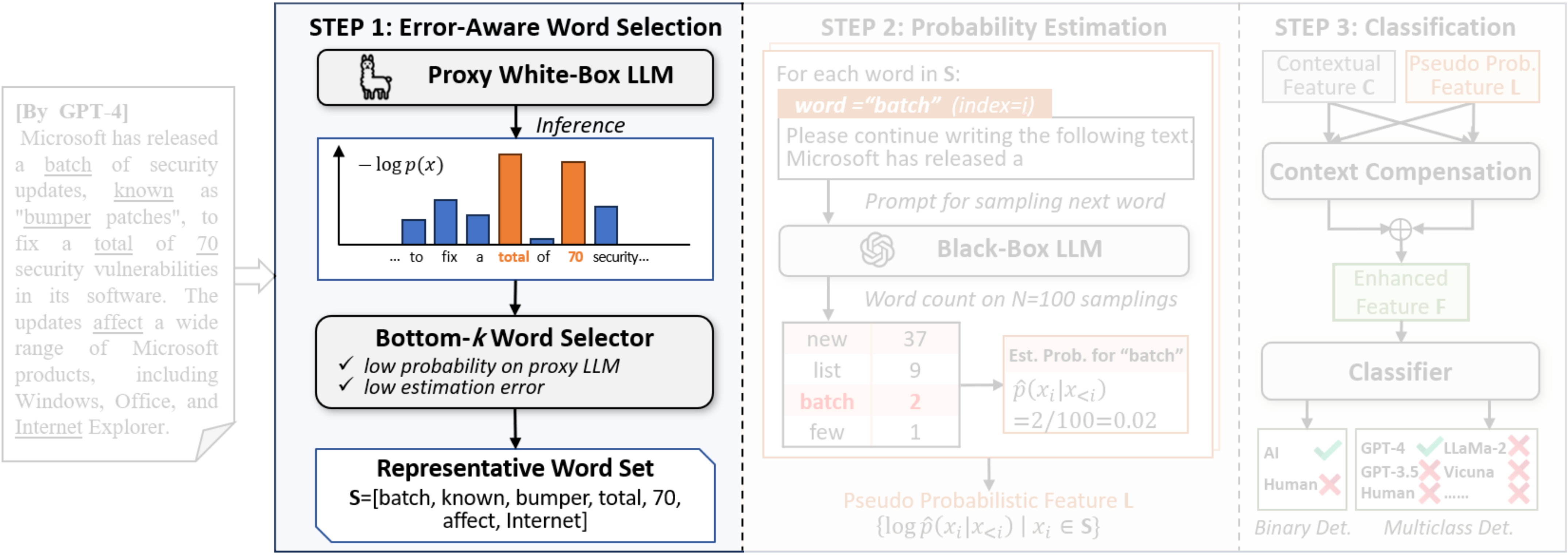
We propose POGER, a **pro**xy-**g**uided **e**fficient **r**e-sampling method.



Step 1: Error-Aware Word Selection

- Use an easy-to-use LM (e.g., GPT-2) as the proxy to infer on the given text x and obtain token probabilities $\boldsymbol{p}^\theta = (p_1^\theta, p_2^\theta, \cdots, p_n^\theta)$
- Adopt an error-aware bottom-k word selector to get the representative word set S:

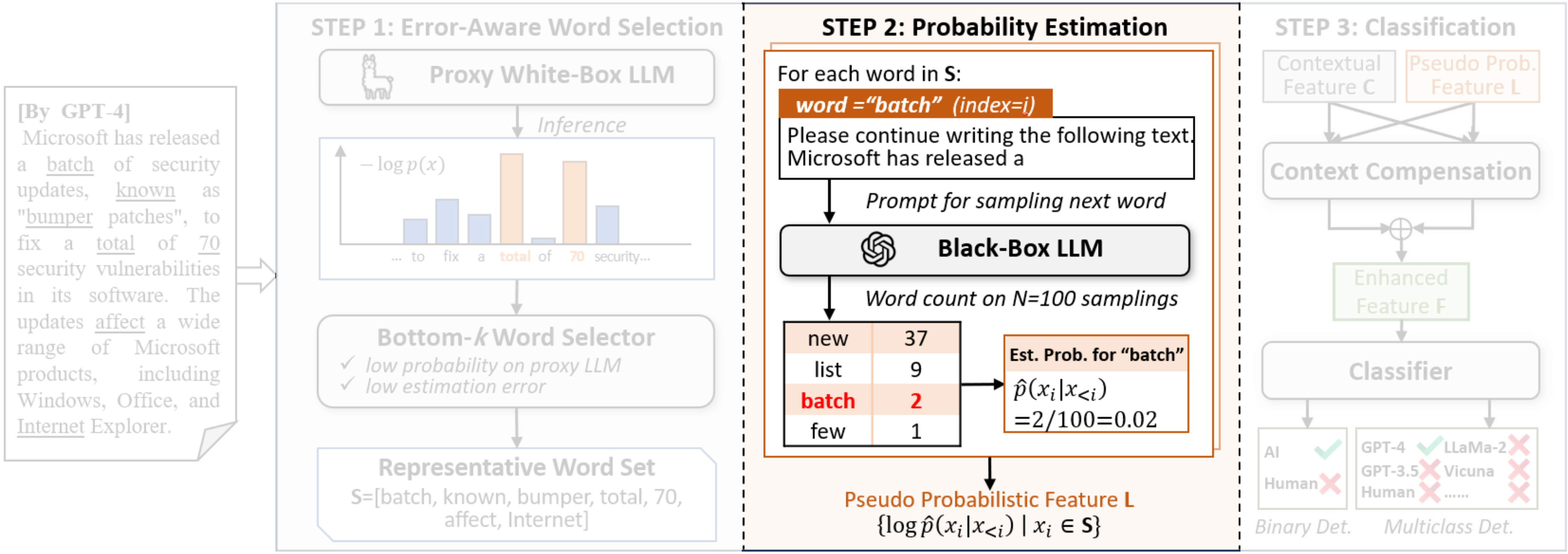
$$\boldsymbol{p}^{\theta\prime} = \left\{ p_i \middle| p_i \geq \frac{1}{1+N\Delta^2} \right\} \qquad \mathrm{IDX} = \left\{ i \middle| p_i^\theta \in \mathrm{MINK}(\boldsymbol{p}^{\theta\prime}) \right\}, \mathrm{S} = \{ x_i | i \in \mathrm{IDX} \}$$



Step 2: Probability Estimation

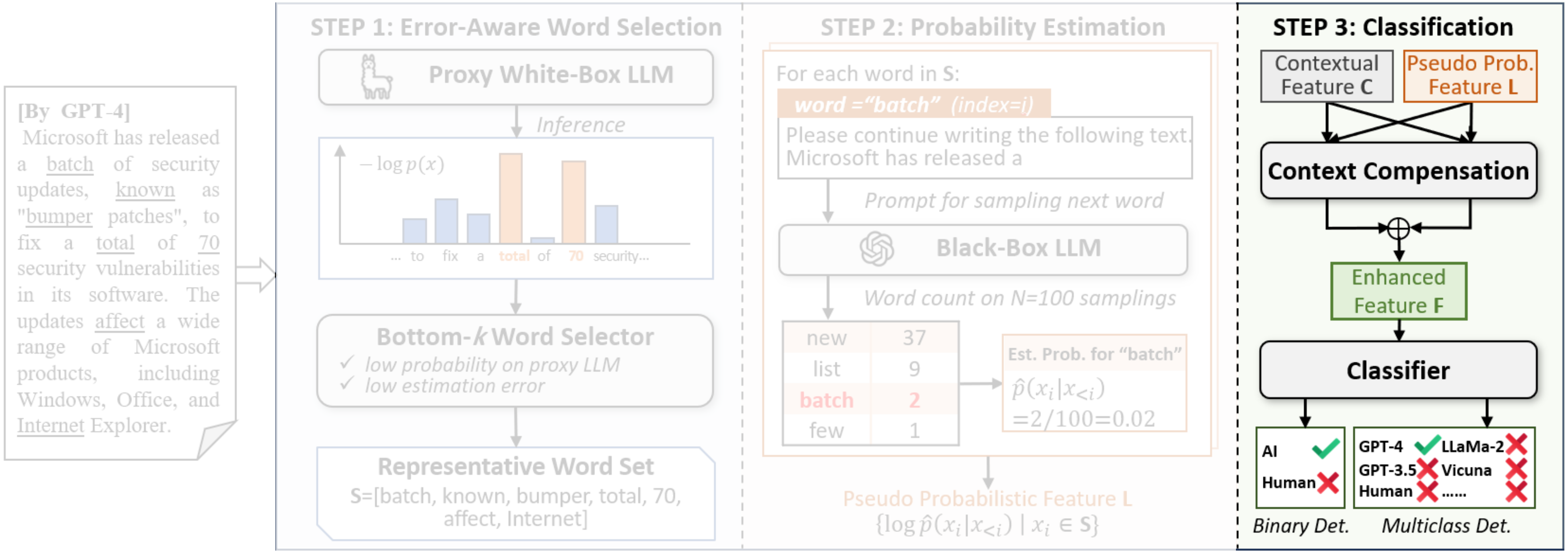
- Sample and calculate probability for the selected k words in S on the given M candidate black-box LLMs by N times
- Get the pseudo log probabilistic feature matrix:

$$\mathbf{L} = [\boldsymbol{l}_i]_{i=1}^{k} \in \mathbb{R}^{k \times M} \qquad \boldsymbol{l}_i = \left[\hat{p}_{\theta_j}\left(x_{\mathrm{IDX}[i]} \middle| x_{\mathrm{IDX}[i]-b:\mathrm{IDX}[i]-1}\right)\right]_{j=1}^{M}$$



Step 3: Context-Compensated Classification

- As context compensation, introduce the contextual semantic representation $\mathbf{C} \in \mathbb{R}^{k \times d}$
- $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V}$, $\mathbf{F} = \text{Attention}(\mathbf{L}, \mathbf{C}, \mathbf{C}) \oplus \text{Attention}(\mathbf{C}, \mathbf{L}, \mathbf{L})$
- $\hat{y} = \text{softmax}(\text{MLP}(\mathbf{F}))$



Evaluation

Method	Human	GPT-2	GPT-J	LLaMA-2	Vicuna	Alpaca	GPT-3.5	GPT-4	MacF1
Partial White-Box Setting									
DNA-GPT White	N/A	62.70	40.79	45.36	30.49	70.18	N/A	N/A	49.91*
Sniffer	96.60	**100.00**	**100.00**	98.49	95.85	**99.23**	75.34	72.65	92.27
SeqXGPT	**98.07**	**100.00**	99.62	**98.88**	**99.62**	98.87	85.93	84.17	95.64
POGER-Mixture	97.32	98.88	99.23	98.11	97.71	98.86	**97.36**	**97.38**	**98.11**
w/o Context Compensation	96.97	99.62	99.23	96.68	94.94	98.48	95.42	95.13	97.06
Black-Box Setting									
RoBERTa	88.24	78.03	86.55	55.47	58.70	59.91	70.63	84.13	72.71
T5-Sentinel	87.29	85.42	88.71	67.78	62.11	69.73	75.79	79.83	77.08
DNA-GPT Black	N/A	38.58	21.56	48.80	33.85	47.15	53.99	39.82	40.53*
Sniffer	87.41	89.82	87.26	29.52	47.62	35.84	34.21	52.63	58.04
SeqXGPT	91.67	89.66	86.77	23.64	46.31	45.64	42.10	62.40	61.02
POGER	**92.49**	**93.75**	**89.96**	**90.49**	**89.30**	**93.82**	**90.98**	**92.59**	**91.67**
w/o Context Compensation	84.21	88.30	80.63	81.88	88.65	91.95	89.49	87.35	86.56

- POGER outperforms all baseline methods in both settings of multiclass AIGT detection.
- POGER has better OOD generalization capabilities, benefiting from the pseudo probabilistic.

Method	In-Dist.	Out-of-Distribution			
		QA→Writing		Writing→QA	
RoBERTa	72.71	54.23	*(-25.42%)*	46.73	*(-35.73%)*
T5-Sentinel	77.08	47.23	*(-38.73%)*	53.19	*(-30.99%)*
Sniffer	58.04	57.50	*(-0.93%)*	53.16	*(-8.41%)*
SeqXGPT	61.02	59.07	*(-3.20%)*	54.94	*(-9.96%)*
POGER	**91.67**	**89.00**	*(-2.91%)*	**84.19**	*(-8.16%)*

Conclusion

- **Motivation:** Achieve "white-boxizing" the black-box LLM by estimating word generation probability through multiple re-sampling, so that the high performance white-box detection methods can also be used under black-box setting.
- **Method:** By selecting low-probability words as representative words, the number of re-samples can be significantly reduced, thus improving efficiency and reducing costs.
- **Result:** Experiments on texts from humans and 7 LLMs demonstrated the superiority of POGER.

Data & Code

https://github.com/ICTMCG/POGER

Paper List

https://github.com/ICTMCG/Awesome-Machine-Generated-Text

Our Latest Work

(WeChat, in chinese)

THANKS

shiyuhui22s@ict.ac.cn