ICLR

Glimpse: Enabling White-Box Methods to Use Proprietary Models for Zero-Shot LLM-Generated Text Detection

Guangsheng Bao, Yanbin Zhao, Juncai He, Yue Zhang
(baoguangsheng@westlake.edu.cn)

Zhejiang University

Westlake University

Shanghai Polytechnic University

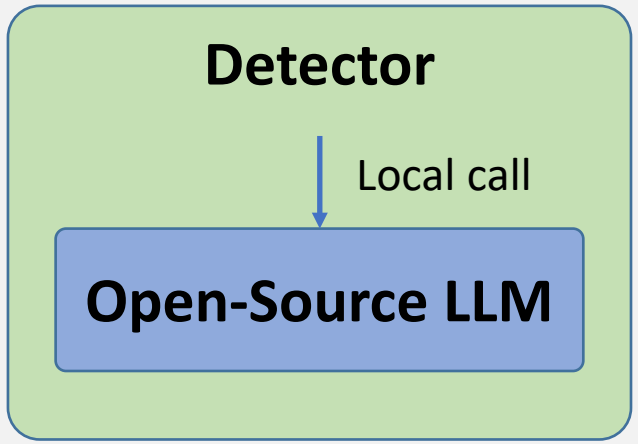
King Abdullah University of Science and Technology

Motivation

西湖大學 WESTLAKE UNIVERSITY

Metric-Based Detectors

White-Box Methods

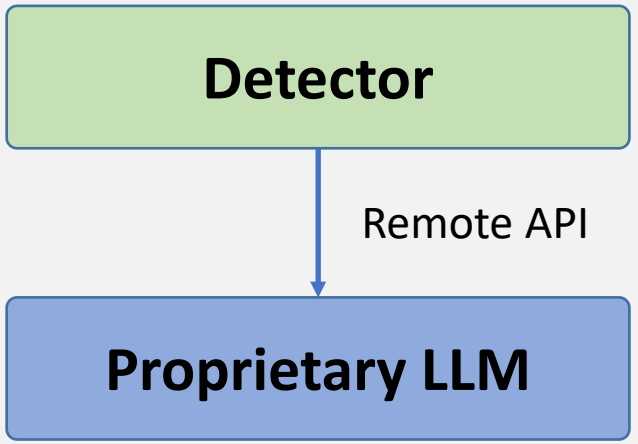


Characteristics:

- Weaker model
- Quick and low cost
- High detection accuracy

vs

Black-Box Methods



Characteristics:

- Stronger model
- Slow and high cost
- Low detection accuracy

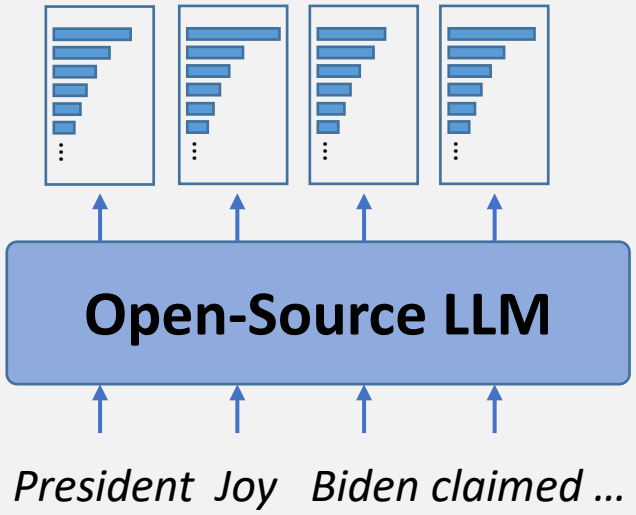
Motivation

西湖大學
WESTLAKE UNIVERSITY

Challenge for Combination

White-Box Methods

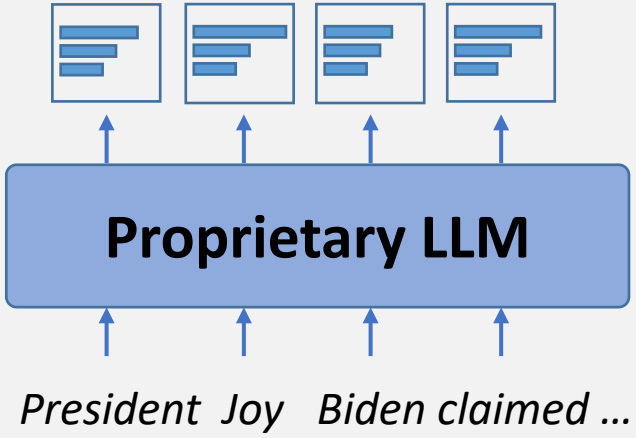
(full distribution)



vs

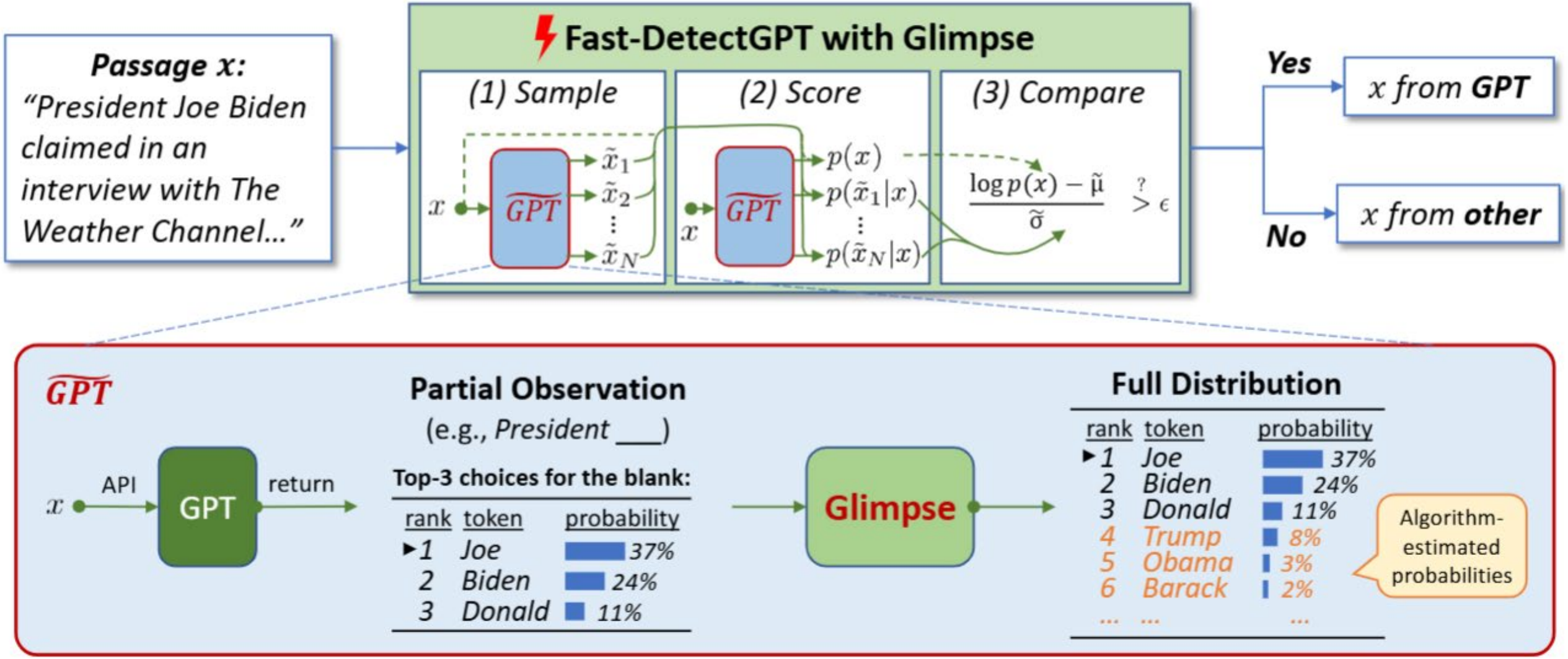
API-based Proprietary Models

(partial observation)



Our Method

西湖大學
WESTLAKE UNIVERSITY



Experiments

西湖大學
WESTLAKE UNIVERSITY

Overview

Method	ChatGPT	GPT-4	Claude-3 Sonnet	Claude-3 Opus	Gemini-1.5 Pro	Avg.
Fast-DetectGPT *(Open-Source: GPT-Neo-2.7B)*	0.9487	0.8999	0.9260	0.9468	0.8072	0.9057
Glimpse (Fast-DetectGPT) *(Proprietary: GPT-3.5)*	**0.9766** (↑ **54%**)	**0.9411** (↑ **41%**)	**0.9576** (↑ **43%**)	**0.9689** (↑ **42%**)	**0.9244** (↑ **61%**)	**0.9537** (↑ **51%**)

Experiments

西湖大學 WESTLAKE UNIVERSITY

Multilingual

Method	Chinese (Web QA)	Russian (RuATD)	Urdu (News)	Indonesian (News)	Arabic (Wikipedia)	Bulgarian (News)	Mix6
Fast-Detect (GPT-J/Neo-2.7B)	0.9319	0.8158	0.9630	0.9876	0.9121	0.9422	0.8862
Fast-Detect (Phi2-2.7B)	0.8024	0.6710	0.9049	0.9313	0.9155	0.7500	0.8205
Fast-Detect (Qwen2.5-7B)	0.5345	0.7065	0.9970	0.9687	0.8404	0.8992	0.8063
Fast-Detect (Llama3-8B)	0.8729	0.7984	0.9962	0.9922	0.9282	0.9723	0.8713
Glimpse using Geometric							
Fast-Detect (Babbage)	0.9814	0.7673	0.9894	0.9857	0.9916	0.9628	0.8950
Fast-Detect (Davinci)	**0.9938**	0.8030	0.9996	0.9991	0.9982	0.9842	0.9369
Fast-Detect (GPT-3.5)	0.9913	0.8555	**1.0000**	0.9996	**0.9999**	**0.9925**	**0.9774**

Q & A

西湖大學
WESTLAKE UNIVERSITY

Thank you!

Check our GitHub for the demo, code, and paper.