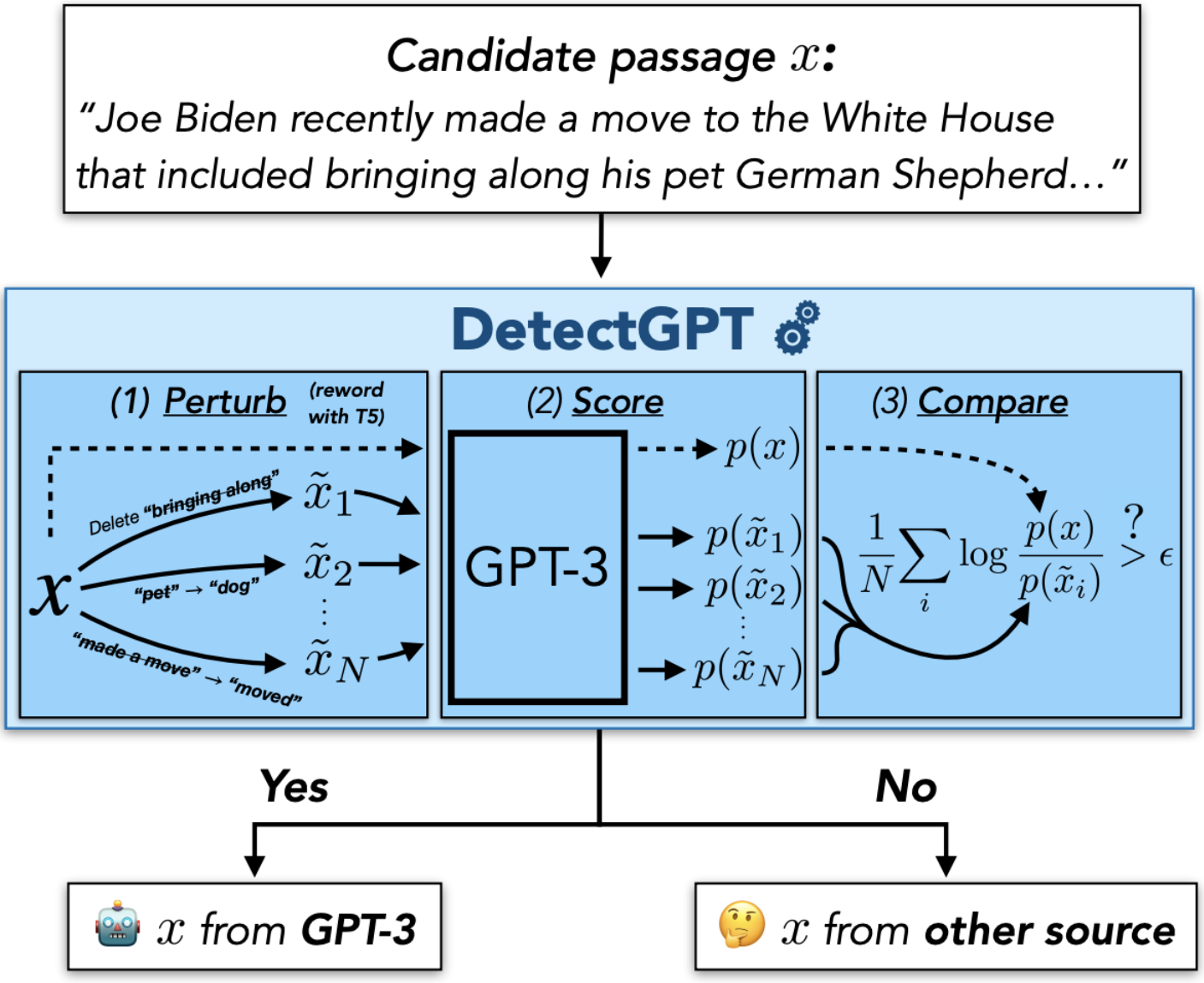
Efficient Detection of LLM-generated Texts with a Bayesian Surrogate Model(ACL 2024)

Kwan Ho Lee

August 4, 2025

Seoul National University

DetectGPT


Candidate passage x:
"Joe Biden recently made a move to the White House that included bringing along his pet German Shepherd…"
DetectGPT
(1) Perturb (reword with T5)
(2) Score
(3) Compare
Delete "bringing along"
"pet" → "dog"
"made a move" → "moved"
x
$\tilde{x}_1$
$\tilde{x}_2$
$\tilde{x}_N$
GPT-3
$p(x)$
$p(\tilde{x}_1)$
$p(\tilde{x}_2)$
$p(\tilde{x}_N)$
$\frac{1}{N}\sum_i \log \frac{p(x)}{p(\tilde{x}_i)} \overset{?}{>} \epsilon$
Yes
No
x from GPT-3
x from other source

Surrogate Model

- When an outcome of interest cannot be easily measured or computed, an approximate mathematical model of the outcome is used instead.
- In our case, we will fit a Gaussian Process model with the text as x and $y = log(p(x))$. In other words, we aim to build a model that replaces steps (1) through (2) outlined above.

Bayesian Surrogate Model

Fitting Gaussian Process Regression

- Kernel function $\mathcal{K}(x, x') = \alpha \cdot \mathrm{BERTScore}(x, x') + \beta$
- Train Data
  - Make pertubation text $\{x_i\}_{i=1}^{N} = \{x_1\} \cup \mathrm{X}^*$
  - train with $\mathrm{X}_t = \{x_0(\text{machine-generated text}) \ , \ x_1\}$
  - retrain with $X_{t+1} = \mathrm{X}_t \cup x \qquad : x \ = \ \arg\max_{x \in X^*} \sigma_t^2(x)$
- Posterior Equations

$$p(f_{x^*} \mid \mathrm{X}_t, \mathrm{y}_t, \mathrm{X}^*) = \mathcal{N}(\bar{f}^*, \Sigma^*)$$

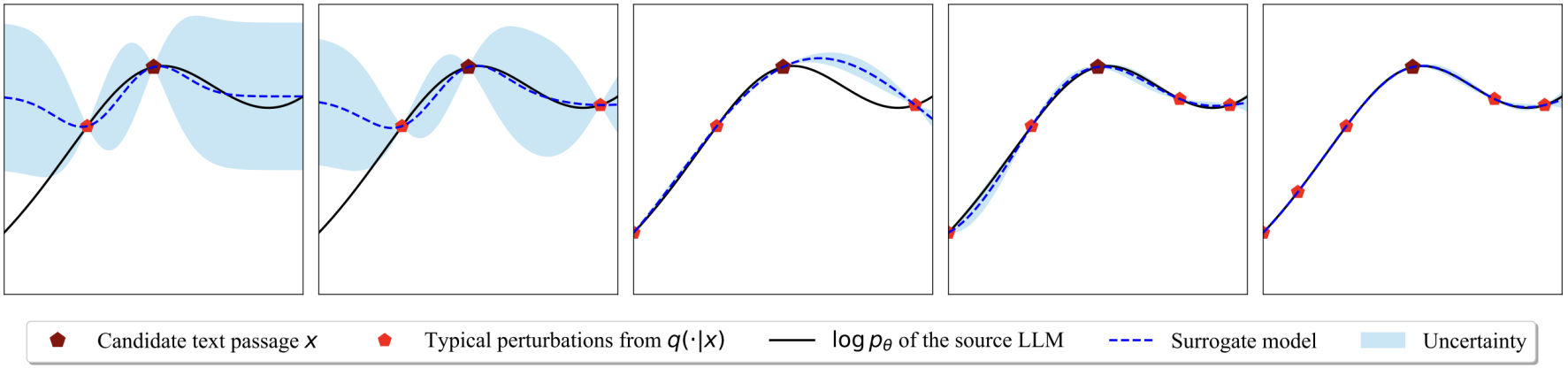
where

$$f(x) \sim \mathcal{GP}\left(0, k(x, x')\right) \tag{1}$$

$$\bar{f}^* := k_{\mathrm{x}^*, \mathrm{x}_t} \left[k_{\mathrm{x}_t, \mathrm{x}_t} + \sigma^2 \mathrm{I}\right]^{-1} \mathrm{y}_t \tag{2}$$

$$\Sigma^* := k_{\mathrm{x}^*, \mathrm{x}^*} - k_{\mathrm{x}^*, \mathrm{x}_t} \left[k_{\mathrm{x}_t, \mathrm{x}_t} + \sigma^2 \mathrm{I}\right]^{-1} k_{\mathrm{x}_t, \mathrm{x}^*} \tag{3}$$

Bayesian Surrogate Model



- Why choose a Gaussian Process as the surrogate model?
  - Performs well in low-data regimes (only a few queried points).
  - Resistance for overfitting
  - Capability to quantify uncertainty
  - Flexible kernels capture non-trivial local curvature.
- Query efficiency: This method detects a single machine-generated text with just 4 queries, whereas DetectGPT requires about 100 for the same task.

Experiment

Experiment

Experimenting with different text-generation models

Dataset	Method	2	3	4	10	20	50	100	200
Xsum	DetectGPT (7B)	0.850	0.893	0.904	0.927	0.932	0.95	0.952	0.952
	Our Method (7B)	0.958	0.955	0.957	-	-	-	-	-
	DetectGPT (13B)	0.817	0.849	0.861	0.913	0.922	0.936	0.932	0.935
	Our Method (13B)	0.886	0.912	0.929	-	-	-	-	-
Squad	DetectGPT (7B)	0.798	0.820	0.857	0.871	0.878	0.889	0.886	0.884
	Our Method (7B)	0.947	0.936	0.932	-	-	-	-	-
	DetectGPT (13B)	0.671	0.686	0.712	0.740	0.731	0.743	0.758	0.757
	Our Method (13B)	0.799	0.785	0.787	-	-	-	-	-
Writing	DetectGPT (7B)	0.903	0.916	0.930	0.938	0.937	0.936	0.941	0.938
	Our Method (7B)	0.985	0.983	0.979	-	-	-	-	-
	DetectGPT (13B)	0.856	0.897	0.933	0.962	0.963	0.963	0.966	0.967
	Our Method (13B)	0.929	0.971	0.979	-	-	-	-	-

Figure 1: The AUROC for detecting samples generated by Vicuna-7B/13B varies depending on the number of queries made to the source model.

Experiment

White Box Assumption

Method	Xsum	Squad	Writing
DetectGPT (LLaMA2)	0.397	0.473	0.641
Our Method (LLaMA2)	0.631	0.660	0.742
DetectGPT (Vicuna)	0.595	0.606	0.733
Our Method (Vicuna)	0.780	0.714	0.850

Figure 2: The AUROC for detecting texts generated by ChatGPT with query budget 15

Appendix

Appendix - Number of Typical Samples

The Visualization Typical Samples

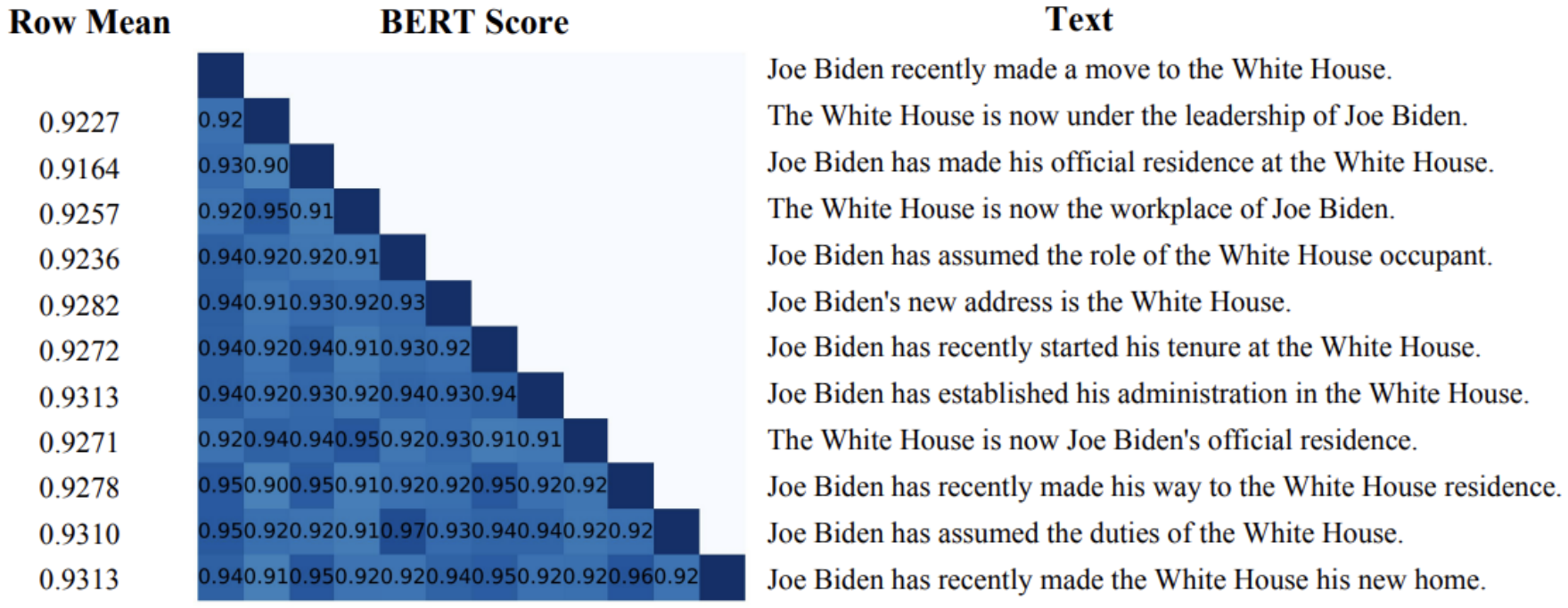


Figure 3: The visualization of the candidate text passage and the first 11 typical perturbations of it identified by our method, ordered from top to bottom. The BertScore among them and the row mean (estimated without the diagonal elements) are reported. The log probabilities are given by GPT-2.

References

[1] Miao, Y., Gao, H., Zhang, H., Deng, Z. (2023). Efficient detection of LLM-generated texts with a Bayesian surrogate model. arXiv preprint arXiv:2305.16617.

[2] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., Finn, C. (2023, July). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In International conference on machine learning (pp. 24950-24962). PMLR.