CSCE 689: Special Topics in Trustworthy NLP

Lecture 14: AI-Generated Text Detection (2)

Kuan-Hao Huang
khhuang@tamu.edu

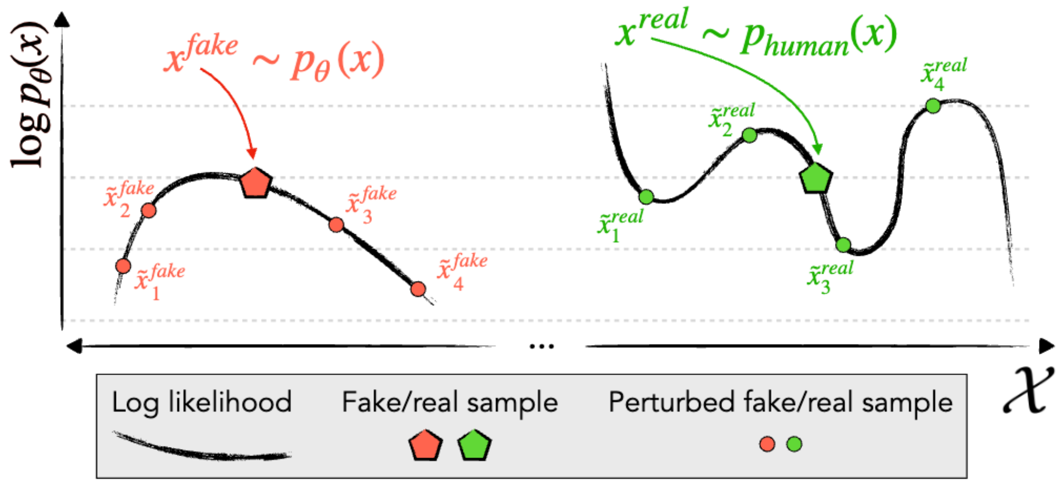
Course Project – Proposal

- Due: 9/25
- Page limit: 2 pages (exclude references)
- Format: ACL style
- The proposal should include
  - The topic you choose
  - An introduction to the task
  - Evaluation metrics
  - The dataset, models, and approaches you plan to use

Recap: Probability Curvature

$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(x)$$

- Should be relatively large when example x is machine-generated
- Should be relatively small when example x is human-written



Conditional Probability Curvature

$$\mathbf{d}(x, p_\theta, q_\varphi) = \frac{\log p_\theta(x|x) - \tilde{\mu}}{\tilde{\sigma}}$$

$$\tilde{\mu} = \mathbb{E}_{\tilde{x} \sim q_\varphi(\tilde{x}|x)} \left[\log p_\theta(\tilde{x}|x)\right] \quad \textbf{and} \quad \tilde{\sigma}^2 = \mathbb{E}_{\tilde{x} \sim q_\varphi(\tilde{x}|x)} \left[(\log p_\theta(\tilde{x}|x) - \tilde{\mu})^2\right]$$

Probability curvature proposed by DetectGPT

$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x} \sim q(\cdot|x)} \log p_\theta(x)$$

Red Teaming Language Model Detectors with Language Models

Zhouxing Shi*, Yihan Wang*, Fan Yin*, Xiangning Chen, Kai-Wei Chang, Cho-Jui Hsieh
University of California, Los Angeles
{zshi, yihanwang, fanyin20, xiangning, kwchang, chohsieh}@cs.ucla.edu
*Alphabetical order

Detectors Can Be Attacked

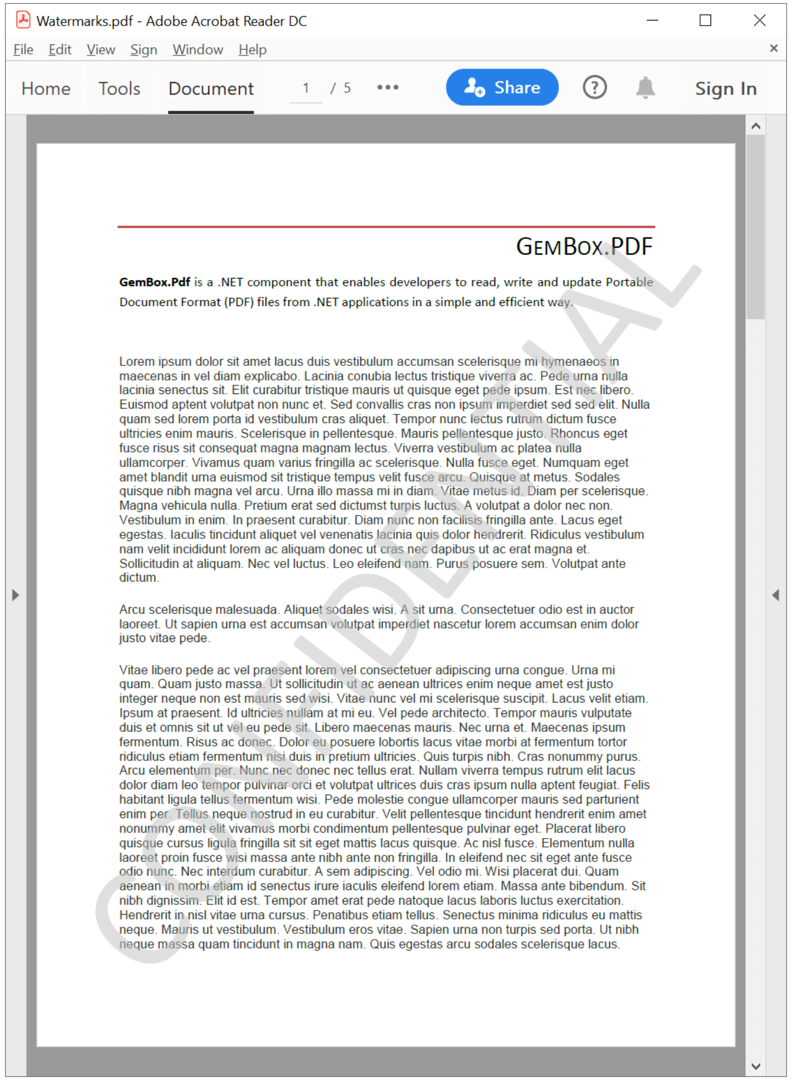
- Perturb machine-generated text
  - **Query-free** word replacement
  - **Query-based** word replacement
  - **Paraphrasing** text

Results

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**

Watermarking

- Post-detection can be hard
- Add watermark during training/generating
  - Watermark should not affect too much to the generation quality
  - Watermark cannot be too obvious
  - Watermark verification needs to be viable
  - Watermark cannot be removed easily



A Watermark for Large Language Models

John Kirchenbauer [*] **Jonas Geiping** [*] **Yuxin Wen** **Jonathan Katz** **Ian Miers** **Tom Goldstein**
University of Maryland

Assumptions

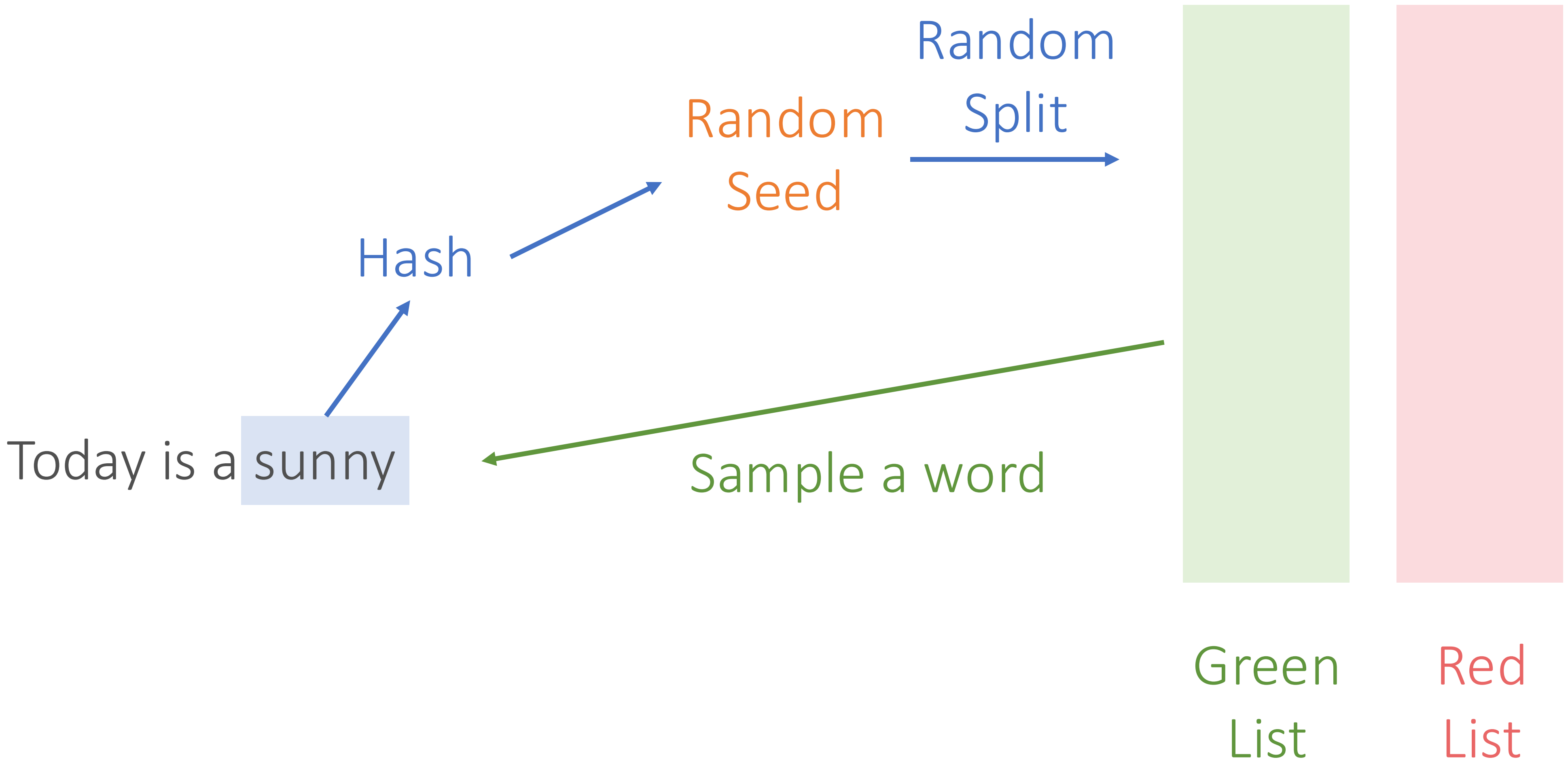
- Add watermark when generating texts
- We have the access to the vocabulary of the model

Watermarking Example

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

How to decide green/red words?

Text Generation with Hard Red List



Text Generation with Hard Red List

- The chance of a random text has a valid watermark
  - $\left(\frac{1}{2}\right)^{T}$ for a length T text
- Watermark detection
  - Statistic way: one proportion z-test

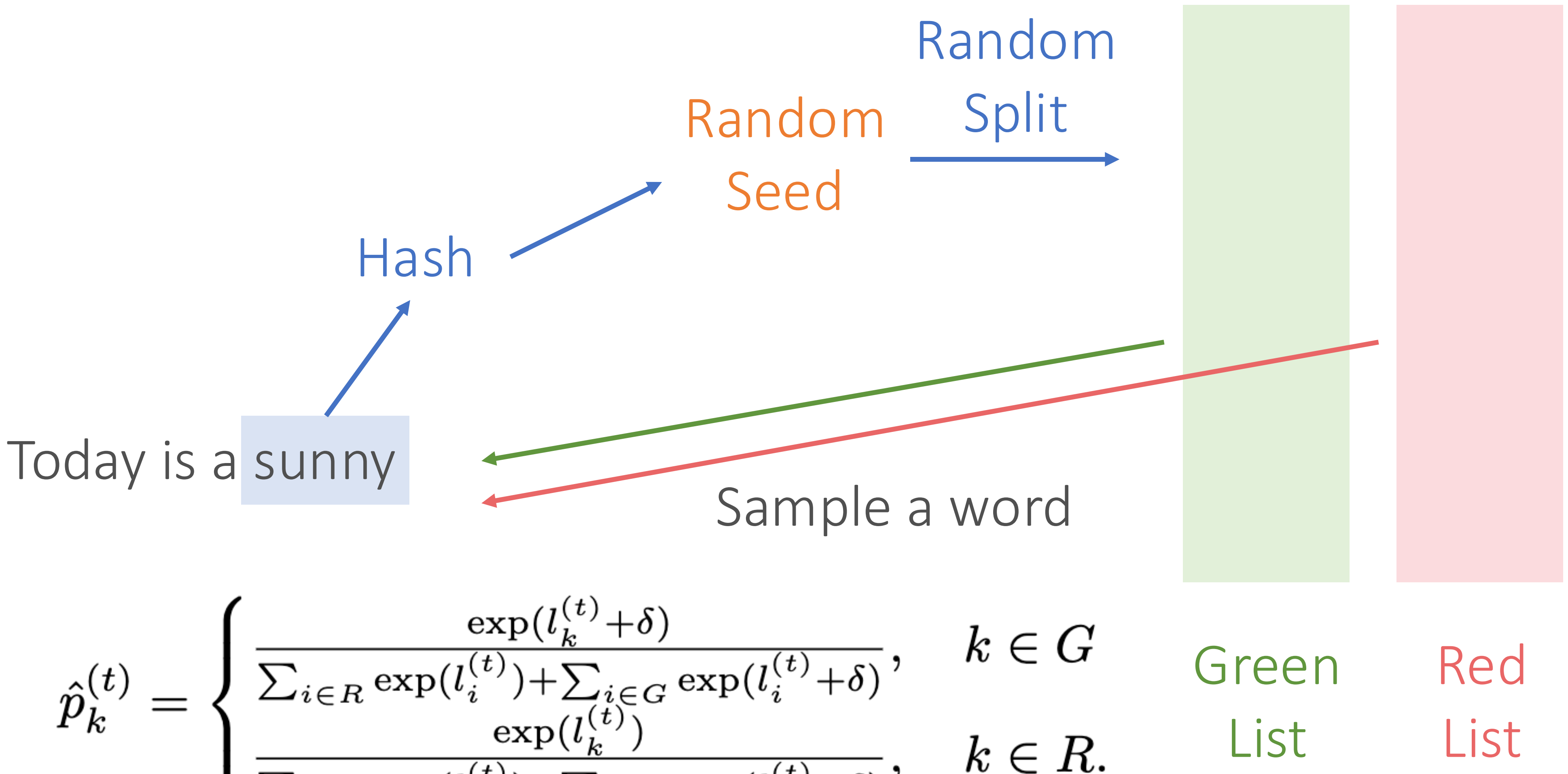
$$z = 2(|s|_G - T/2)/\sqrt{T}.$$

  - If z > threshold → having watermark
  - z > 4, the probability of a false positive is 3×10e-5

Text Generation with Hard Red List

- Generated texts can be not natural for certain cases
  - Barack Obama

Text Generation with Soft Red List



$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)}+\delta)}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in R. \end{cases}$$

Text Generation with Soft Red List

Algorithm 2 Text Generation with Soft Red List

Input: prompt, $s^{(-N_p)} \cdots s^{(-1)}$
green list size, $\gamma \in (0, 1)$
hardness parameter, $\delta > 0$

for $t = 0, 1, \cdots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \cdots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.

2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.

3. Using this random number generator, randomly partition the vocabulary into a "green list" G of size $\gamma|V|$, and a "red list" R of size $(1-\gamma)|V|$.

4. Add δ to each green list logit. Apply the softmax operator to these modified logits to get a probability distribution over the vocabulary.

$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)}+\delta)}{\sum_{i \in R} \exp(l_i^{(t)})+\sum_{i \in G} \exp(l_i^{(t)}+\delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i \in R} \exp(l_i^{(t)})+\sum_{i \in G} \exp(l_i^{(t)}+\delta)}, & k \in R. \end{cases}$$

5. Sample the next token, $s^{(t)}$, using the watermarked distribution $\hat{p}^{(t)}$.

end for

Text Generation with Soft Red List

Theorem 4.2. *Consider watermarked text sequences of T tokens. Each sequence is produced by sequentially sampling a raw probability vector $p^{(t)}$ from the language model, sampling a random green list of size γN, and boosting the green list logits by δ using Equation 4 before sampling each token. Define $\alpha = \exp(\delta)$, and let $|s|_G$ denote the number of green list tokens in sequence s.*

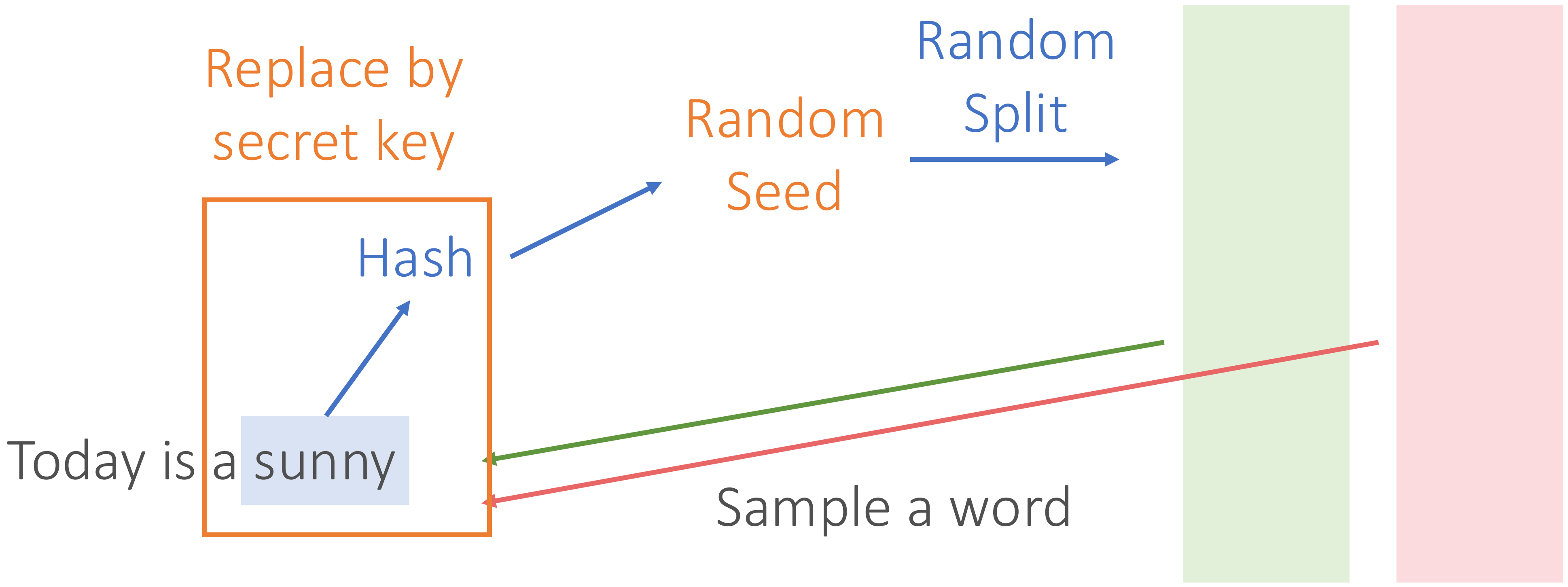
If a randomly generated watermarked sequence has average spike entropy at least $S^\star$, i.e.,

$$\frac{1}{T}\sum_t S\left(p^{(t)}, \frac{(1-\gamma)(\alpha-1)}{1+(\alpha-1)\gamma}\right) \geq S^\star,$$

then the number of green list tokens in the sequence has expected value at least

$$\mathbb{E}\,|s|_G \geq \frac{\gamma\alpha T}{1+(\alpha-1)\gamma}S^\star,$$

Private Watermarking

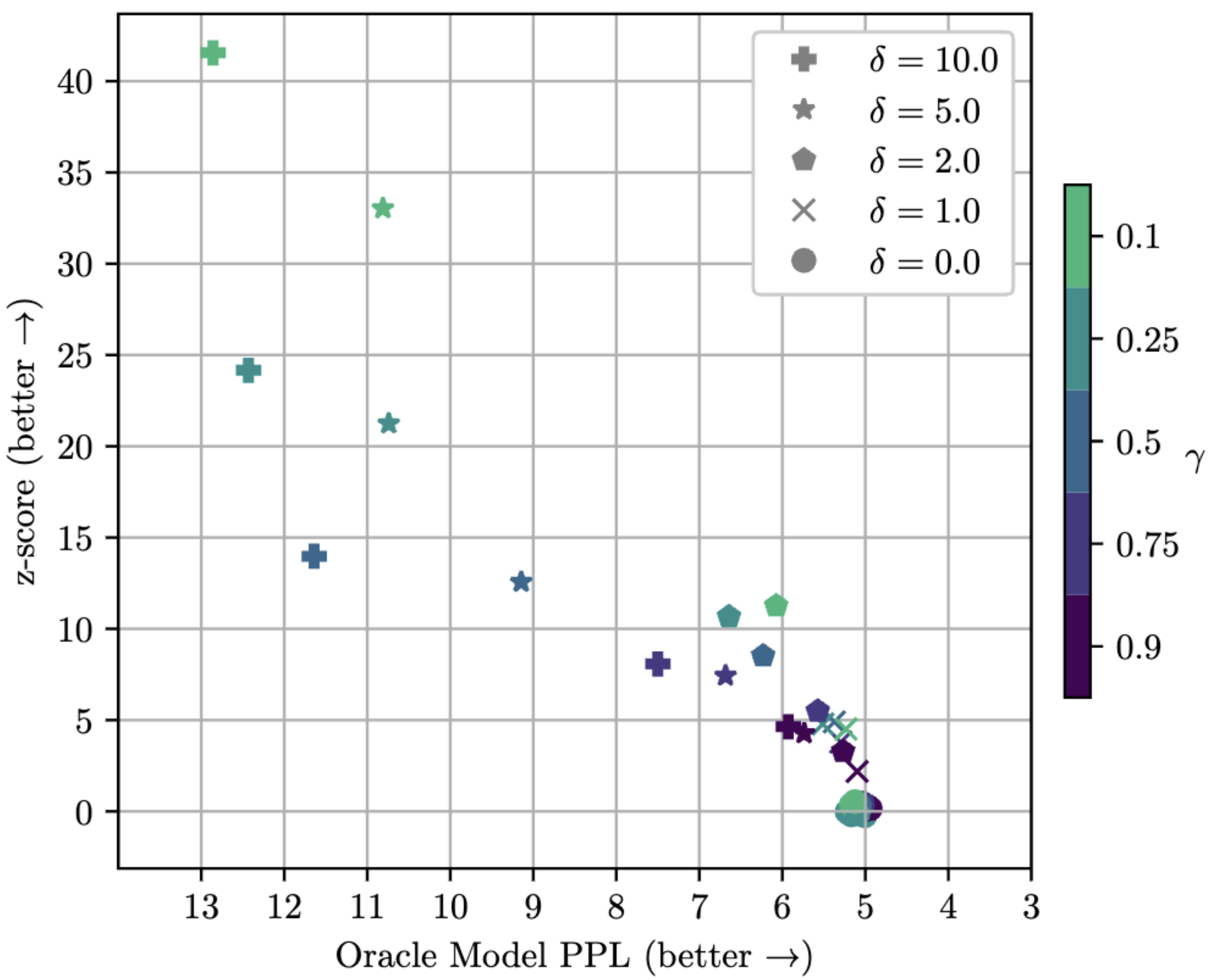


$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)}+\delta)}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)}+\delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)}+\delta)}, & k \in R. \end{cases}$$

Green List

Red List

Text Generation Quality



Text Generation Quality

prompt	real completion	no watermark (NW)	watermarked (W)
...tled out of court and publicly reconciled.\nIn the '80s the band's popularity waned in the United States but remained strong abroad. Robin released three solo albums, with limited success. The Bee Gees	returned with some moderate hits in the late 1990s and were inducted into the Rock and Roll Hall of Fame in 1997. With his brothers, Mr. Gibb won six Grammys.\nIn addition to his wife and his brother [...continues]	continued to tour, and Barry became a television producer.\nBut in the early '90s, the Bee Gees' popularity remained high. They scored a hit with "Don't Stop Believing" in 1990, and in 1992 the Bee Ge[...continues]	' 1990 album, "Spirits of the Century," was a mixed critical and commercial success.\nWhen the brothers were nominated for a Grammy Award in 1990, Mr. Gibb's "You Should Be Dancing" and "Massachusetts,[...continues]
... logged into their Google account and have verified profiles that match queries for the site.\nGoogle's John Mueller said there is no ranking benefit in using different Google Search Console and Google	Analytics accounts for each individual web site you manage. The topic came up before, as long as you are not spamming Google - there also is no down side to using the same accounts across multiple we[...continues]	Analytics data to calculate ranking sentiment.\nGoogle have a new feature called Keyword Difficulty Analysis that shows you the keywords your competitors are ranking for. It shows the demand curve as [...continues]	+ accounts to see different numbers.\nGoogle also released their own great blog post on the news algorithm. They give lots of great advice to help your site do better.\nFinally, at the end of September [...continues]

Watermark Detection Results

				z=4.0				z=5.0			
sampling	δ	γ	count	FPR	TNR	TPR	FNR	FPR	TNR	TPR	FNR
m-nom.	1.0	0.50	506	0.0	1.0	0.767	0.233	0.0	1.0	0.504	0.496
m-nom.	1.0	0.25	506	0.0	1.0	0.729	0.271	0.0	1.0	0.482	0.518
m-nom.	2.0	0.50	507	0.0	1.0	0.984	0.016	0.0	1.0	0.978	0.022
m-nom.	2.0	0.25	505	0.0	1.0	0.994	0.006	0.0	1.0	0.988	0.012
m-nom.	5.0	0.50	504	0.0	1.0	0.996	0.004	0.0	1.0	0.992	0.008
m-nom.	5.0	0.25	503	0.0	1.0	1.000	0.000	0.0	1.0	0.998	0.002
8-beams	1.0	0.50	495	0.0	1.0	0.873	0.127	0.0	1.0	0.812	0.188
8-beams	1.0	0.25	496	0.0	1.0	0.819	0.181	0.0	1.0	0.770	0.230
8-beams	2.0	0.50	496	0.0	1.0	0.992	0.008	0.0	1.0	0.984	0.016
8-beams	2.0	0.25	496	0.0	1.0	0.994	0.006	0.0	1.0	0.990	0.010
8-beams	5.0	0.50	496	0.0	1.0	1.000	0.000	0.0	1.0	1.000	0.000
8-beams	5.0	0.25	496	0.0	1.0	1.000	0.000	0.0	1.0	1.000	0.000

How About Attacks?

- Perturb machine-generated text
  - **Query-free** word replacement
  - **Query-based** word replacement
  - **Paraphrasing** text

Attacking Results

sampling	ε	count	TPR@4.0	FNR@4.0	w/attck TPR@4.0	w/attck FNR@4.0	TPR@5.0	FNR@5.0	w/attck TPR@5.0	w/attck FNR@5.0
m-nom.	0.1	487	0.984	0.016	0.819	0.181	0.977	0.023	0.577	0.423
m-nom.	0.3	487	0.984	0.016	0.353	0.647	0.977	0.023	0.127	0.873
m-nom.	0.5	487	0.984	0.016	0.094	0.906	0.977	0.023	0.029	0.971
m-nom.	0.7	487	0.984	0.016	0.039	0.961	0.977	0.023	0.012	0.988
beams	0.1	489	0.998	0.002	0.834	0.166	0.998	0.002	0.751	0.249
beams	0.3	489	0.998	0.002	0.652	0.348	0.998	0.002	0.521	0.479
beams	0.5	489	0.998	0.002	0.464	0.536	0.998	0.002	0.299	0.701
beams	0.7	489	0.998	0.002	0.299	0.701	0.998	0.002	0.155	0.845

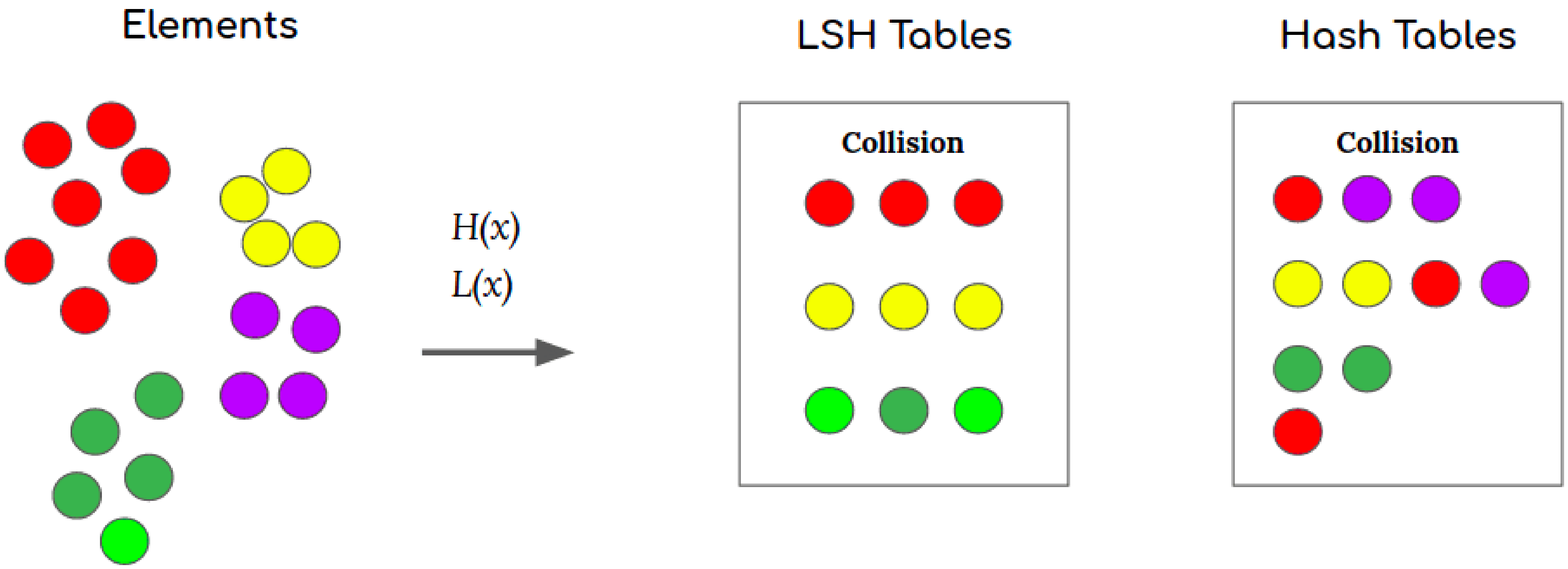
SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation

Abe Bohan Hou[♣*] **Jingyu Zhang**[♣*] **Tianxing He**[♡*]
Yichen Wang[◇] **Yung-Sung Chuang**[♠] **Hongwei Wang**[‡] **Lingfeng Shen**[♣]
Benjamin Van Durme[♣] **Daniel Khashabi**[♣] **Yulia Tsvetkov**[♡]
[♣]Johns Hopkins University [♡]University of Washington [◇]Xi'an Jiaotong University
[♠]Massachusetts Institute of Technology [‡]Tencent AI Lab
{bhou4, jzhan237}@jhu.edu goosehe@cs.washington.edu

How About Attacks?

- Perturb machine-generated text
  - **Query-free** word replacement
  - **Query-based** word replacement
  - **Paraphrasing** text

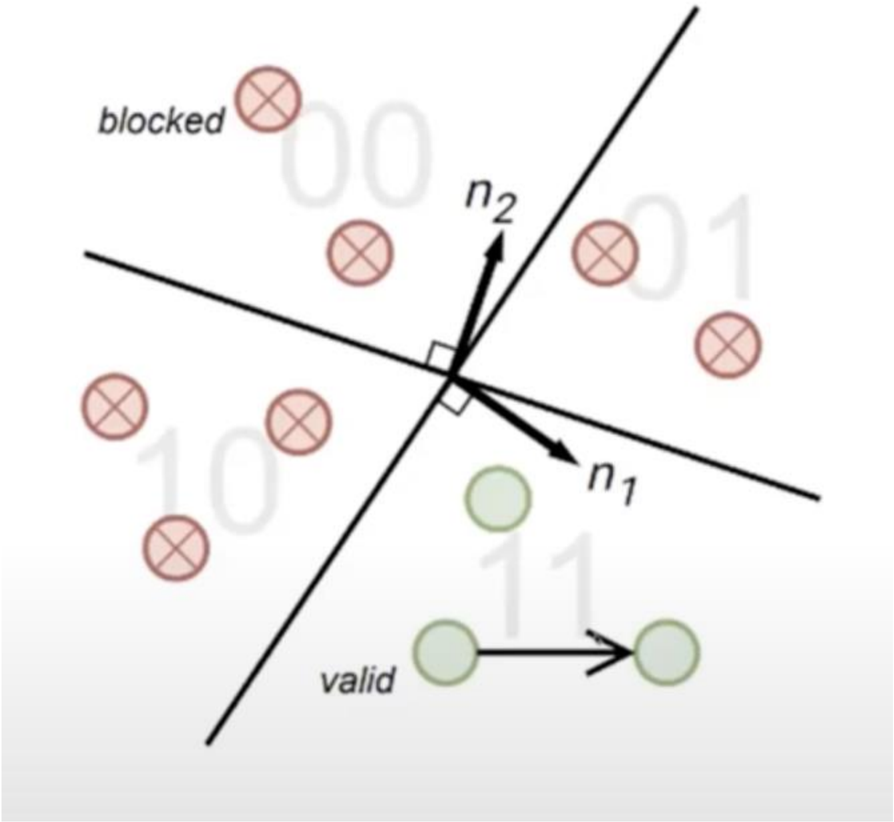
Locality-Sensitive Hashing (LSH)



iunera

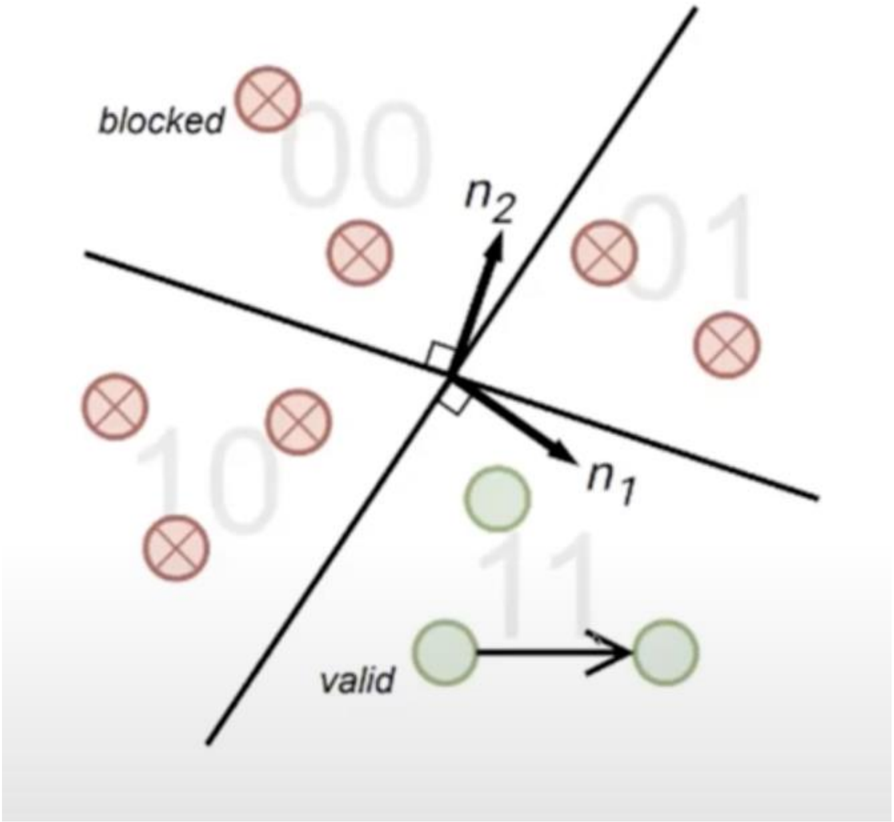
Sentence Encoder

- Semantic encoder robust to paraphrasing
  - SentenceBERT, SimCSE, etc.

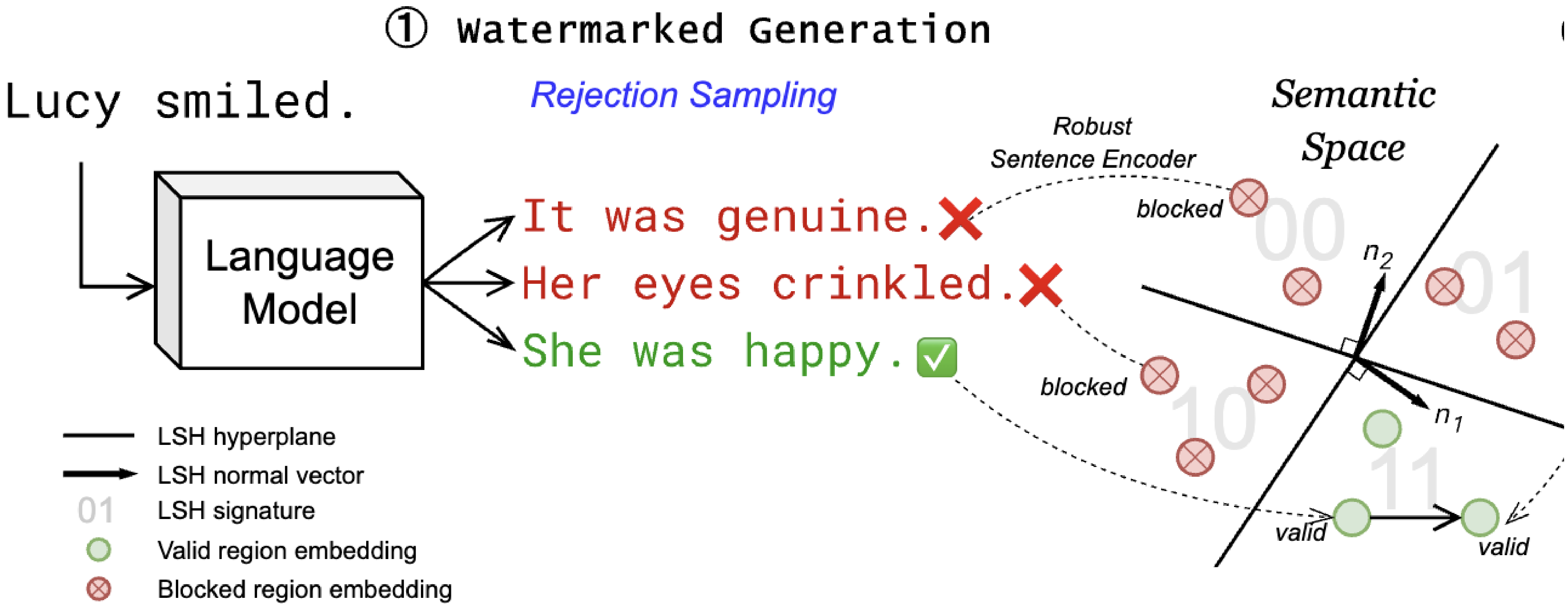


Partition with LSH

- Each dot is a potential next sentence sampled from LM
- LSH partitions the semantic space through random hyperplanes
- Divide the semantic space into valid and blocked regions by hashing on the previous sentence



Generation Overview



Paraphrase Attack

② Paraphrase Attack

Watermark remains valid after paraphrase

✅ She felt delighted.

③ Watermark detecton

No Watermark

Today the company announced results for the third quarter of 2017. The company's board of directors also declared a quarterly cash dividend of $0.23 per share. The dividend is payable to shareholders of record on November 14, 2017. Shareholders are invited to attend the company's annual meeting to propose and discuss a proposal to adopt a new long-term stockholder's plan. The meeting will be held on December 7, 2017.

z-test → human written

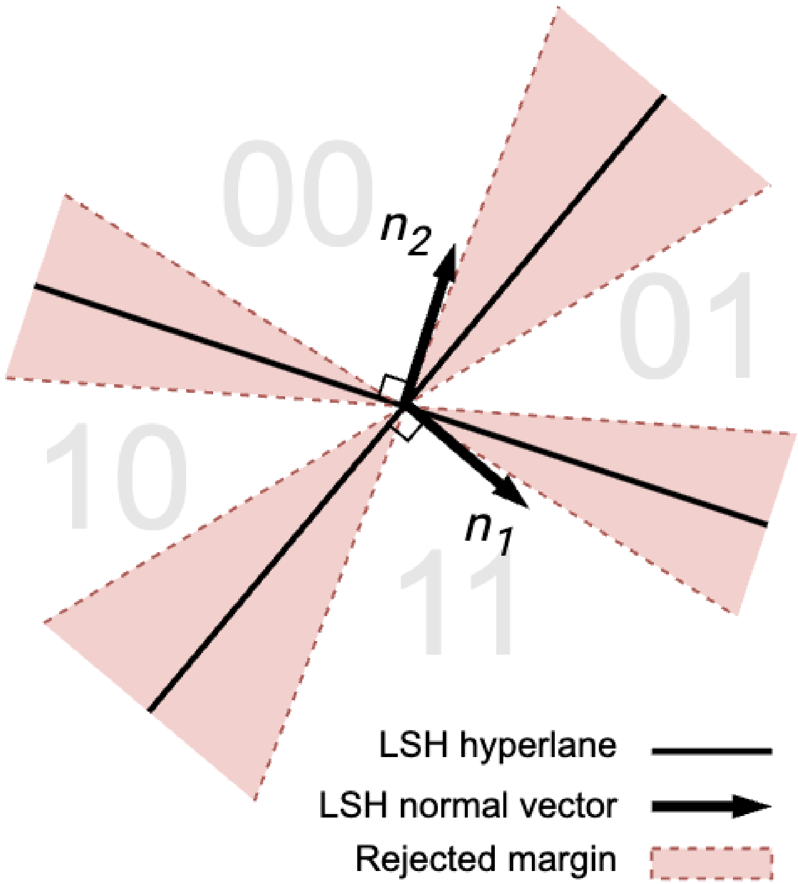
SEMSTAMP

Today the company announced quarterly results for the period ending October 31, 2017. The company also provided an update on its ongoing Phase 3 clinical trial of the Phase 2/3 B-cellderived T cell engager program. These results are included in a newly released Current Report on Form 8-K for the period ending September 30, 2017. You can read the full report at www.curis.com.

z-test → machine written

Consider Margin for Robustness

- Sentence encoder is not perfect
- Only accept sentences with distance larger than a margin



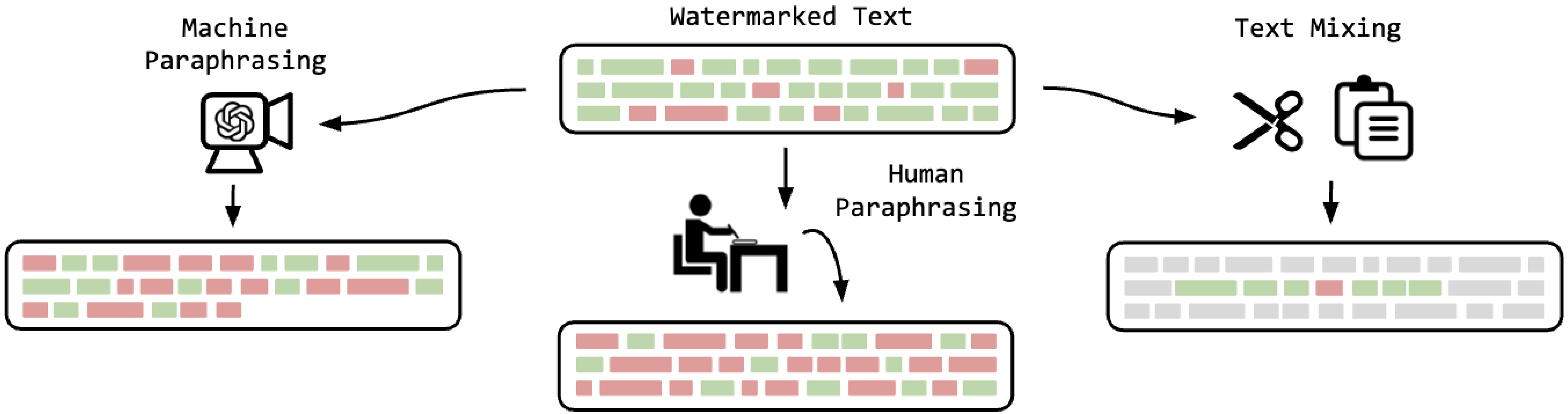
Results

Paraphraser	Algorithm	AUC ↑	TP@1% ↑	TP@5% ↑
		RealNews | BookSum | Reddit-TIFU		
No Paraphrase	KGW	99.6 | 99.9 | 99.3	98.4 | 99.4 | 97.5	98.9 | 99.5 | 98.1
	SSTAMP	99.2 | 99.7 | 99.7	93.9 | 98.8 | 97.7	97.1 | 99.1 | 98.2
Pegasus	KGW	95.9 | 97.3 | 94.1	82.1 | 89.7 | 87.2	91.0 | 95.3 | 87.2
	SSTAMP	**97.8 | 99.2 | 98.4**	**83.7 | 90.1 | 92.8**	**92.0 | 96.8 | 95.4**
Pegasus-bigram	KGW	92.1 | 96.5 | 91.7	42.7 | 56.6 | 67.2	72.9 | 85.3 | 67.6
	SSTAMP	**96.5 | 98.9 | 98.0**	**76.7 | 86.8 | 89.0**	**86.0 | 94.6 | 92.9**
Parrot	KGW	88.5 | 94.6 | 79.5	31.5 | 42.0 | 22.8	55.4 | 75.8 | 43.3
	SSTAMP	**93.3 | 97.5 | 90.2**	**56.2 | 70.3 | 56.2**	**75.5 | 88.5 | 70.5**
Parrot-bigram	KGW	83.0 | 93.1 | 82.8	15.0 | 39.9 | 27.6	37.4 | 71.2 | 49.7
	SSTAMP	**93.1 | 97.5 | 93.9**	**54.4 | 71.4 | 71.8**	**74.0 | 89.4 | 82.3**
GPT3.5	KGW	82.8 | 87.6 | 84.1	17.4 | 17.2 | 27.3	46.7 | 52.1 | 50.9
	SSTAMP	**83.3 | 91.8 | 87.7**	**33.9 | 55.0 | 47.5**	**52.9 | 70.8 | 58.2**
GPT3.5-bigram	KGW	75.1 | 77.1 | 79.8	5.9 | 4.4 | 19.3	26.3 | 27.1 | 41.3
	SSTAMP	**82.2 | 90.5 | 87.4**	**31.3 | 47.4 | 43.8**	**48.7 | 63.6 | 55.9**

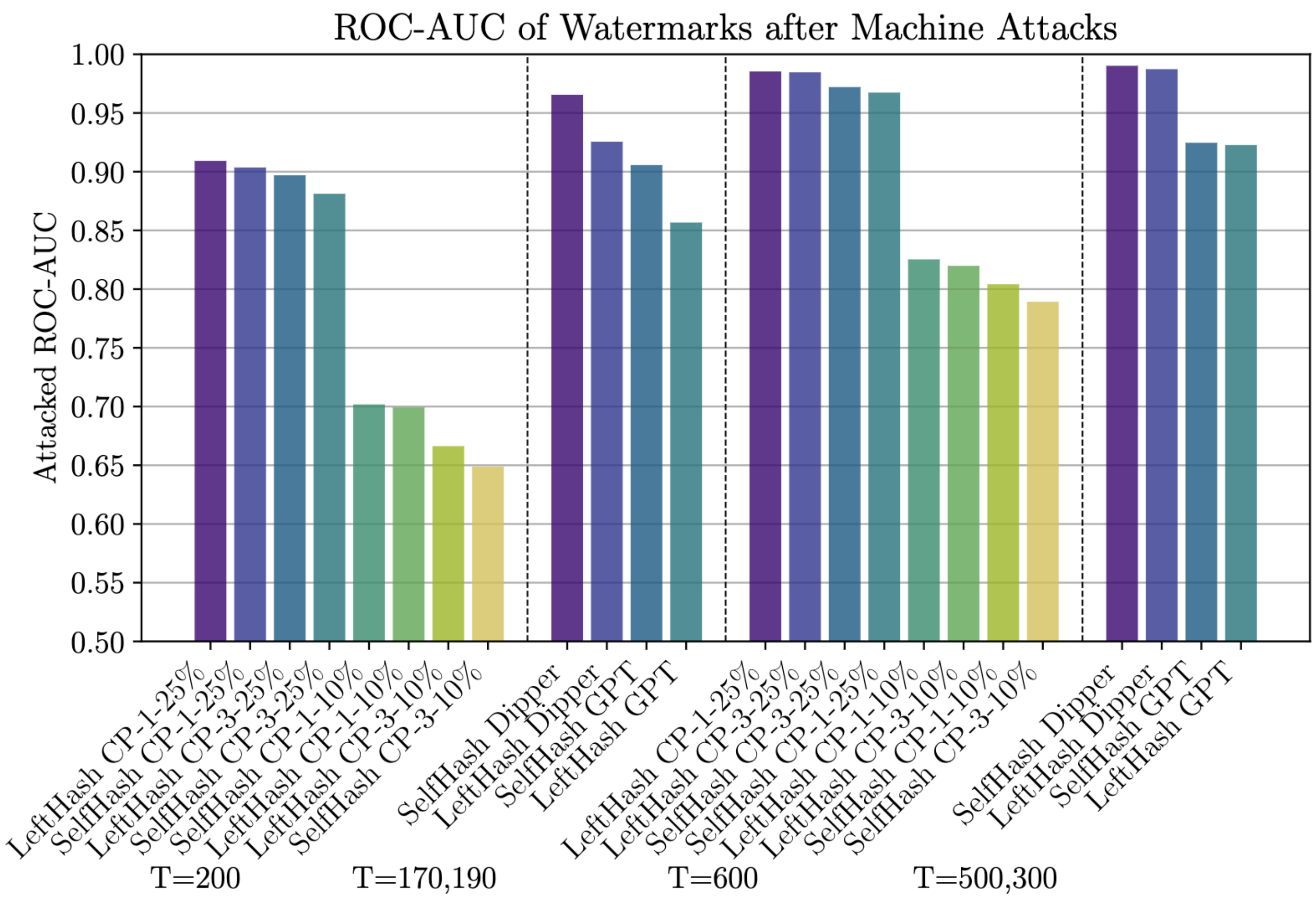
On the Reliability of Watermarks for Large Language Models

John Kirchenbauer[*1], **Jonas Geiping**[*2,3]
Yuxin Wen[1], **Manli Shu**[1], **Khalid Saifullah**[1], **Kezhi Kong**[1],
Kasun Fernando[4], **Aniruddha Saha**[1], **Micah Goldblum**[5], **Tom Goldstein**[1]
[1] University of Maryland
[2] ELLIS Institute Tübingen, [3] Max-Planck Institute for Intelligent Systems, Tübingen AI Center
[4] Scuola Normale Superiore di Pisa, [5] New York University

More Study on Attacks for Token-Level Watermark



Results


ROC-AUC of Watermarks after Machine Attacks
Attacked ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
LeftHash CP-1-25%
SelfHash CP-1-25%
LeftHash CP-3-25%
SelfHash CP-3-25%
SelfHash CP-1-10%
LeftHash CP-1-10%
LeftHash CP-3-10%
SelfHash CP-3-10%
SelfHash Dipper
LeftHash Dipper
SelfHash GPT
LeftHash GPT
LeftHash CP-1-25%
LeftHash CP-3-25%
SelfHash CP-3-25%
SelfHash CP-1-25%
LeftHash CP-1-10%
LeftHash CP-3-10%
SelfHash CP-1-10%
SelfHash CP-3-10%
SelfHash Dipper
LeftHash Dipper
SelfHash GPT
LeftHash GPT
T=200
T=170,190
T=600
T=500,300