UC San Diego

JACOBS SCHOOL OF ENGINEERING

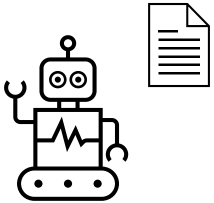
Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models

Mingjia Huo* Sai Ashish Somayajula* Youwei Liang Ruisi Zhang

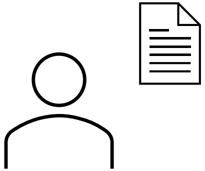
Farinaz Koushanfar Pengtao Xie

University of California, San Diego

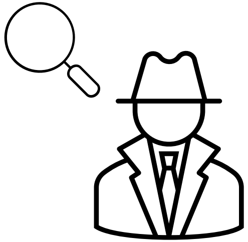
Detecting LLM Generated Texts



LLM generated



Human generated



Detect

Academic dishonesty

Spam content

Misleading content

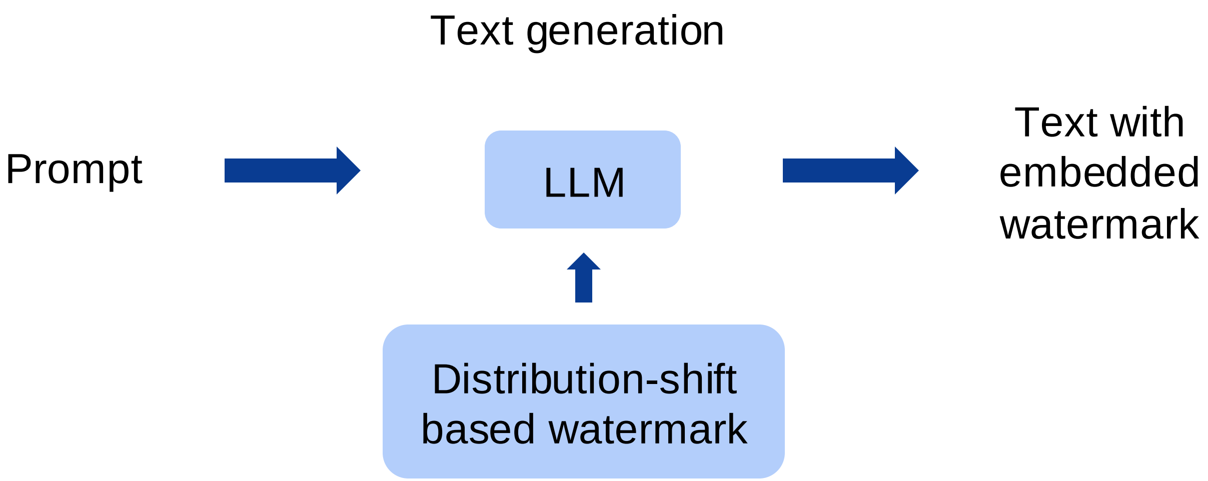
Training degeneration

Prior Methods

Distribution-shift based methods [1, 2, 3]

- Shift the output distribution towards a subset of tokens in the vocabulary
- Statistically estimate the likelihood that the probability distribution has shifted

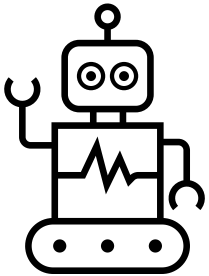
Prior Methods: Distribution-Shift Based Methods



Prior Methods: Distribution-Shift Based Methods

During the generation of t^{th} token,

cat jumped on the |



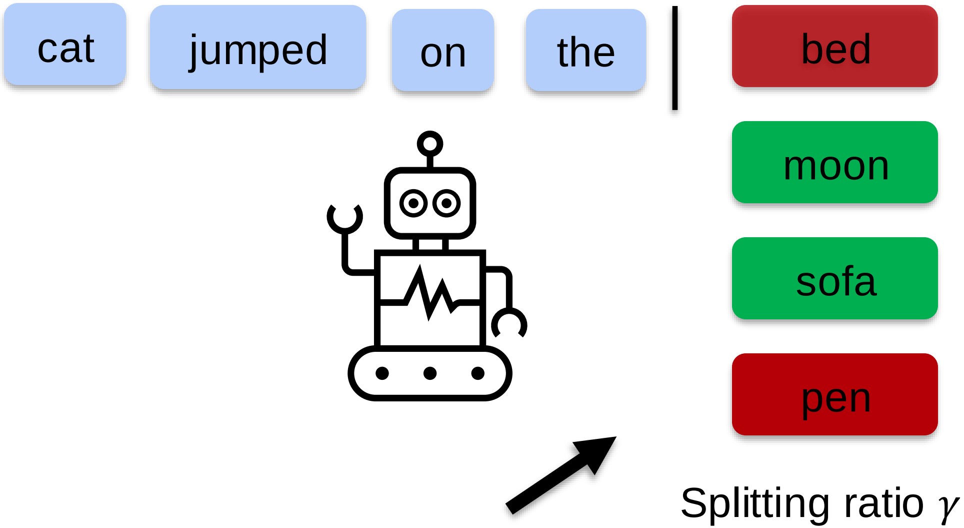
bed

moon

sofa

pen

Prior Methods: Distribution-Shift Based Methods

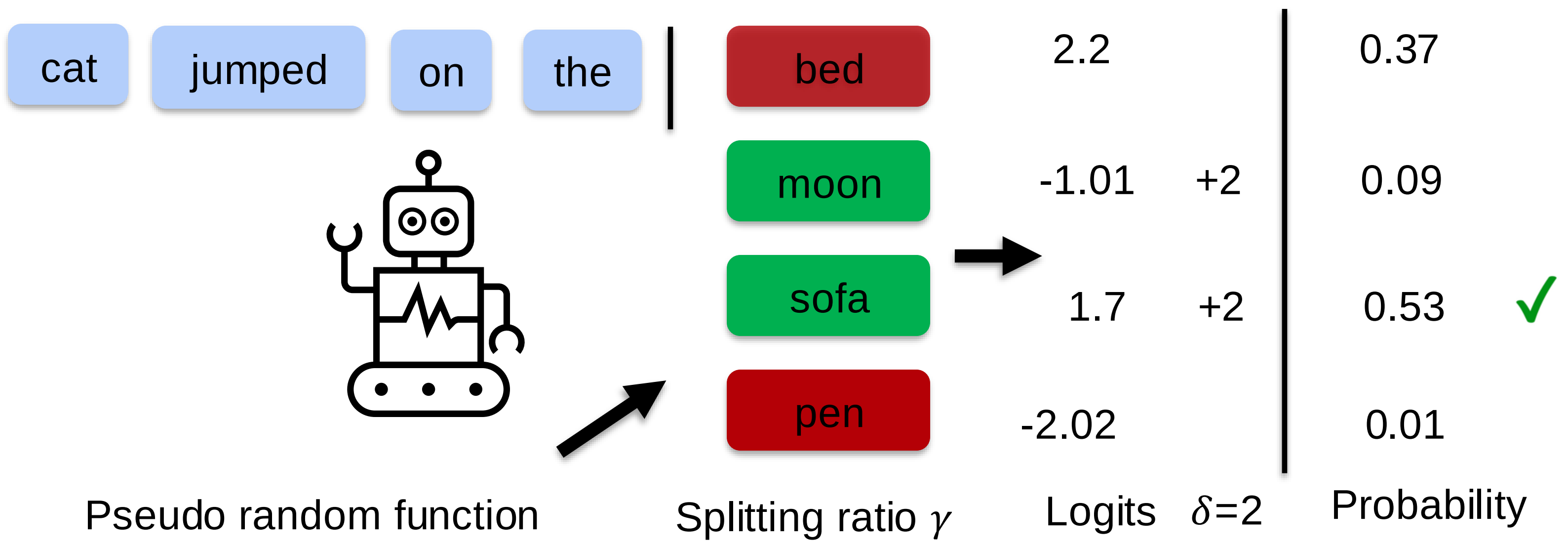


Splitting ratio γ

Pseudo random function

Hash of previous token as seed to partition vocabulary into red-green list

Prior Methods: Distribution-Shift Based Methods

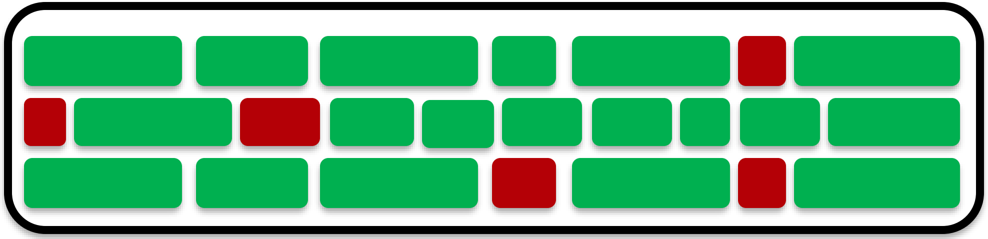


Add δ to all the green tokens to bias the distribution towards green-list

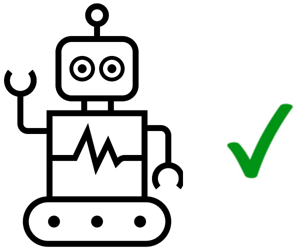
Prior Methods: Distribution-Shift Based Methods

Detection

- Null hypothesis that the next token is selected without the knowledge of green-red list rule, i.e., without addition of δ
- Given hash function, count the number of green tokens in the generation
- Calculate the z-score, $z = \frac{(|s|_G - \gamma T)}{\sqrt{T\gamma(1-\gamma)}}$



Z-score = $\frac{(|s|_G - \gamma T)}{\sqrt{T\gamma(1-\gamma)}}$ = 4



Z-score > τ (say 3)

Limitations

Face challenges in improving the semantics and detectability at the same time

- Improving one compromises the other

Lack adaptive mechanism to adjust γ and δ appropriately

- Ex: Sun rises in the __. It is ‘east’ with certainty. High δ and low γ might not select ‘east’.

Proposed Method

Propose learning token-specific splitting ratio and watermark logit, i.e., γ_t and δ_t

Proposed Method

Propose learning token-specific splitting ratio and watermark logit, i.e., γ_t and δ_t

$s^{(-M)}, \ldots, s^{(-1)}$

Prompt

$s^{(0)}, \ldots, s^{(t-1)}$

Generation till now

Proposed Method

Propose learning token-specific splitting ratio and watermark logit

$s^{(-M)}, \ldots, s^{(-1)}$ $s^{(0)}, \ldots, s^{(t-1)}$ $\longrightarrow$ Embeddings of $s^{(t-1)}$

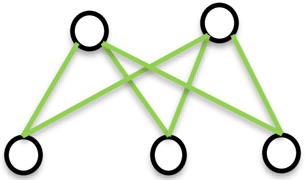
Prompt Generation till now

Proposed Method

Propose learning token-specific splitting ratio and watermark logit

$s^{(-M)}, \dots, s^{(-1)}$ $\quad$ $s^{(0)}, \dots, s^{(t-1)}$ $\longrightarrow$ Embeddings of $s^{(t-1)}$

Prompt $\quad$ Generation till now



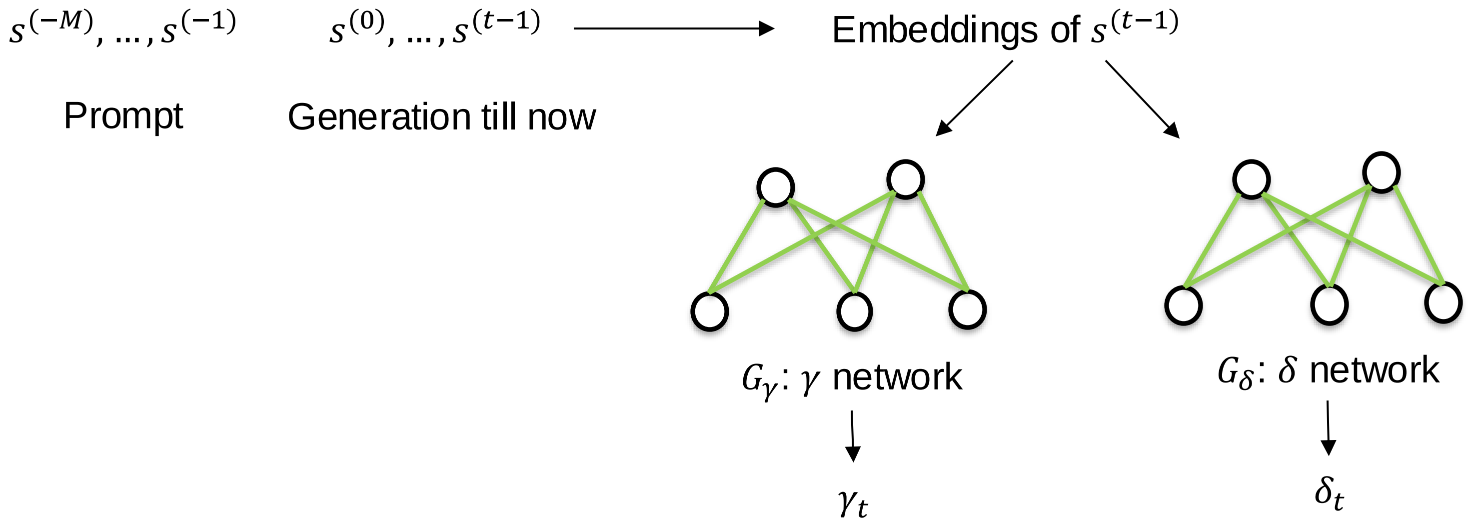
G_γ: γ network

$\downarrow$

γ_t

Proposed Method

Propose learning token-specific splitting ratio and watermark logit



Proposed Method

Differentiable sampling for splitting the vocabulary

- For each token $v \in V$, sample $\mathrm{y}_v^{(t)} \sim B(\gamma_t)$, Bernoulli distribution parameterized by γ_t.
- If $\mathrm{y}_v^{(t)} = 1$, then the token v belongs to green list else red list
- Gumbel softmax trick makes sampling process differentiable

Proposed Method

Given original logits $l_v^{(t)}$ for token v, modified logits after biasing the green-list tokens

$$\hat{\boldsymbol{l}}_v^{(t)} = l_v^{(t)} + y_v^{(t)} * \delta_t$$

Proposed Method

Training objectives

- Detection loss
- Semantic loss

Proposed Method

Detection loss

- Since we have a token-specific γ_t and δ_t, the z-score expression has to be updated based on this distribution

Proposed Method

Theorem: Consider T independent Bernoulli random variables $X_1, \dots, X_T$, each with means $\mu_1, \dots, \mu_T, 0 < \ \mu < 1 \ \forall \ t \in 1, \dots, T$. The sum of these variables, $\sum_{t=1} X_t$, follows a Poisson binomial distribution. When T is sufficiently large, this distribution can be approximated by a Gaussian distribution with mean: $\sum_{t=1}^{T} \mu_t$ and variance: $\sum_{t=1}^{T} \mu_t(1 - \mu_t)$.

Proposed Method

Modified Z-score = $\frac{|s|_G - \sum_{t=1}^{T} \gamma_t}{\sqrt{\sum_{t=1}^{T} \gamma_t (1-\gamma_t)}}$ to account for varying γ_t

Detection loss

- Improve detectability by maximizing this objective
- However, $|s|_G$, count of green tokens, is non-differentiable w.r.t γ_t and δ_t

Proposed Method

Detection loss

- Propose differentiable surrogate $\hat{z} = \frac{\sum_{t=1}^{T} p_{gr}^{(t)} - \sum_{t=1}^{T} \gamma_t}{\sqrt{\sum_{t=1}^{T} \gamma_t (1-\gamma_t)}}$, where $p_{gr}^{(t)}$ is the probability of selecting a green token.
- Maximize $\hat{z}$ or minimize detection loss, $L_D = -\hat{z}$

Proposed Method

Semantic loss

- Generate sentence embeddings of texts before and after watermarking, i.e., s and s_w using the SimCSE model f_θ
- Maximize the cosine similarity between them, $\cos_{sim}(f_\theta(s),\ f_\theta(s_w))$
- Thus, minimize semantic loss, $L_S = -\cos_{sim}(f_\theta(s),\ f_\theta(s_w))$

Proposed Method

Multi-objective Optimization

- Optimizing for two competing loss functions L_D and L_S

  $$\min_{G_\gamma, G_\delta} L_D(G_\gamma, G_\delta) \text{ and } \min_{G_\gamma, G_\delta} L_S(G_\gamma, G_\delta)$$

- Estimate pareto optimal solutions using multiple-gradient descent algorithm (MGDA) [5]

Multiple-Gradient Descent Algorithm

Let g_D and g_S are the gradients of L_D and L_S w.r.t (G_γ, G_δ)

$$\lambda^* = argmin_{\lambda \in [0,1]} \left|\left|\lambda g_D + (1-\lambda) g_S\right|\right|_2$$

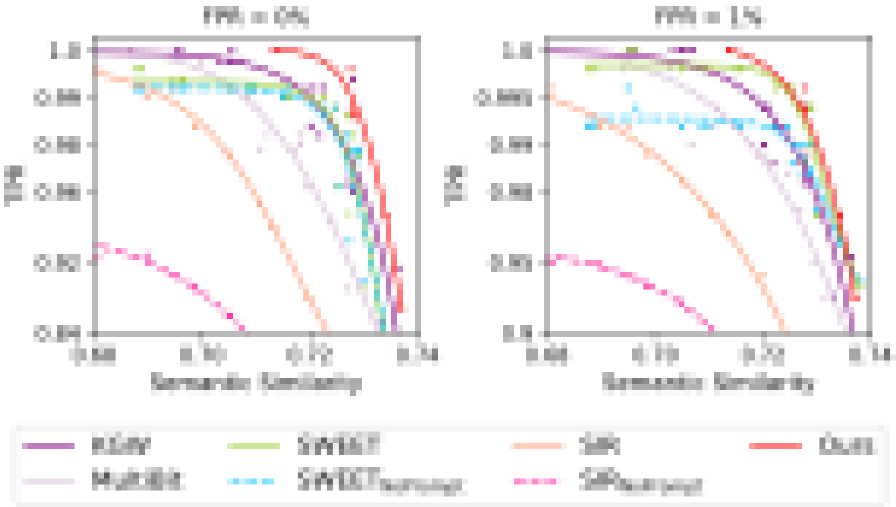
$$g = \lambda^* g_D + (1-\lambda^*) g_S$$

Update (G_γ, G_δ) using the gradient g

Experimental Setup

- Main experiments
  - C4 dataset
    - Training split 6400, Validation split 500, Test split 500
  - Generation length set to 200
- Z-score threshold is empirically determined on respective test sets
  - Set z-score threshold to maintain FPR at 0% and 1%

Results



Comparison of the trade-off for semantic integrity and detectability of different methods applied to OPT-1.3B.

Results

Method	TPR @ 0%	TPR @ 1%	SimCSE
EXP-edit	0.922	0.996	0.655
EXP-edit (Top-k=50)	0.968	0.996	0.677
Ours (Top-k=50)	**1.000**	**1.000**	**0.713**

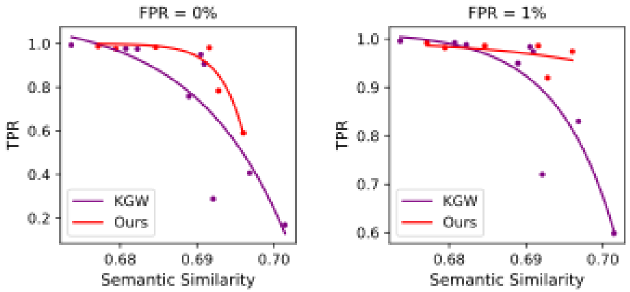
Comparison of our method with indistinguishable method - EXP-edit and its variant EXP-edit (Top-k=50).

Results

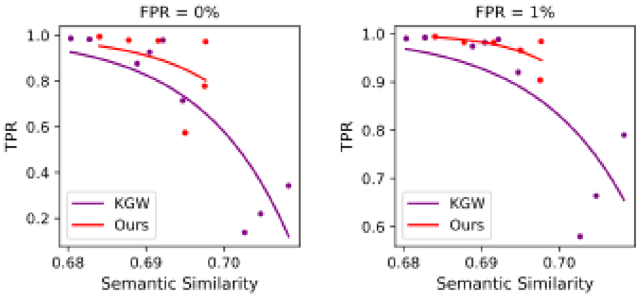
Method	Generation (s)	Detection (s)
No Watermark	3.220	-
KGW	3.827	0.067
SWEET	4.030	0.127
EXP-edit	24.693	155.045
SIR	8.420	0.337
MultiBit	6.500	0.610
Ours	3.946	0.166

Generation and detection speed on OPT-1.3B for generating 200 tokens, measured in seconds.

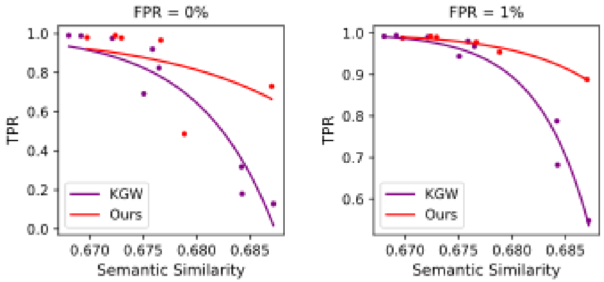
Results



a. LLAMA2 7B



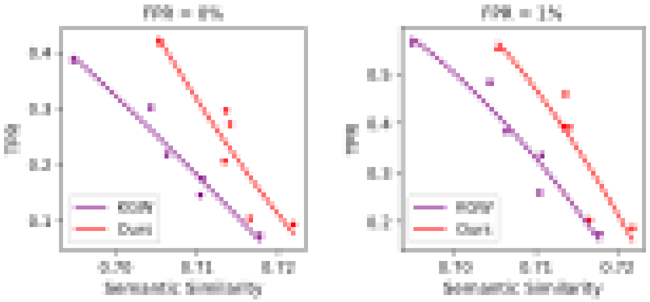
b. LLAMA2 13B



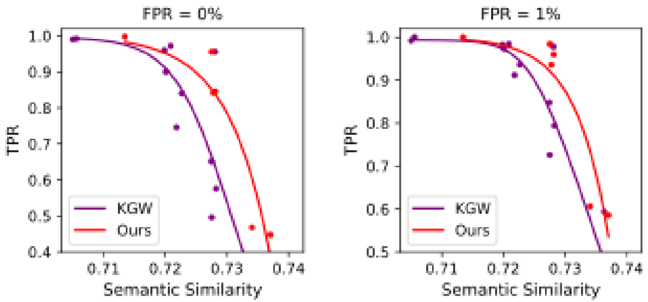
c. LLAMA2 70B

Performance of Ours (trained on OPT-1.3B) and KGW when applied to LLAMA2 7B, 13B, and 70B.

Results



a. Dipper paraphrase attack



b. Copy-Paste-3 attack

Comparison of our method with KGW under dipper paraphrase attack (left) and copy-paste-3 attack (right). Please refer to the paper for other attack results.

Conclusions

- Propose to adapt the watermark strength based on the semantics of the preceding token
- Propose a light-weight network to output token-specific γ_t and δ_t
- Propose a differentiable surrogate of z-score metric for optimization
- Optimize in a multi-objective optimization framework
- Extensive experiments on various scenarios shows the efficacy of our proposed method

References

[1] Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A watermark for large language models." In *International Conference on Machine Learning*, pp. 17061-17084. PMLR, 2023.

[2] Lee, T., Hong, S., Ahn, J., Hong, I., Lee, H., Yun, S., Shin, J., and Kim, G. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*, 2023.

[3] Liu, Aiwei, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. "A semantic invariant robust watermark for large language models." *arXiv preprint arXiv:2310.06356* (2023).

[4] Piet, Julien, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. "Mark my words: Analyzing and evaluating language model watermarks." *arXiv preprint arXiv:2312.00273*(2023).

[5] Sener, Ozan, and Vladlen Koltun. "Multi-task learning as multi-objective optimization." *Advances in neural information processing systems* 31 (2018).