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

UC San Diego JACOBS SCHOOL OF ENGINEERING

Mingjia Huo* Sai Ashish Somayajula* Youwei Liang Ruisi Zhang Farinaz Koushanfar Pengtao Xie

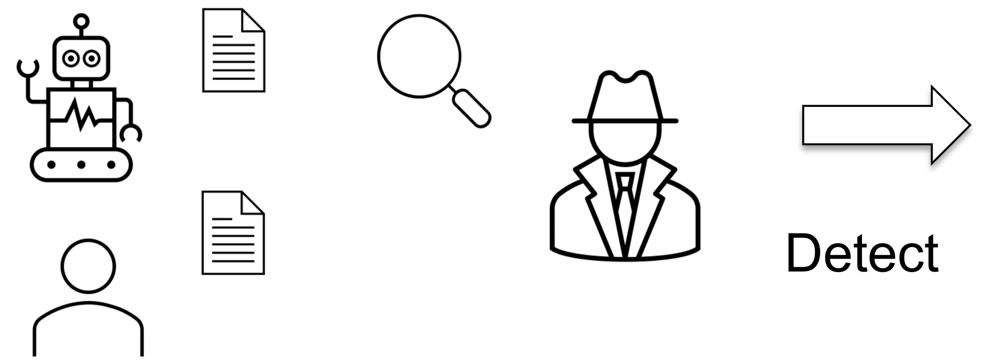
Department of Electrical and Computer Engineering, University of California, San Diego

* Denotes equal contribution

ICML International Conference On Machine Learning

Motivation

Detect between human and machine generated texts



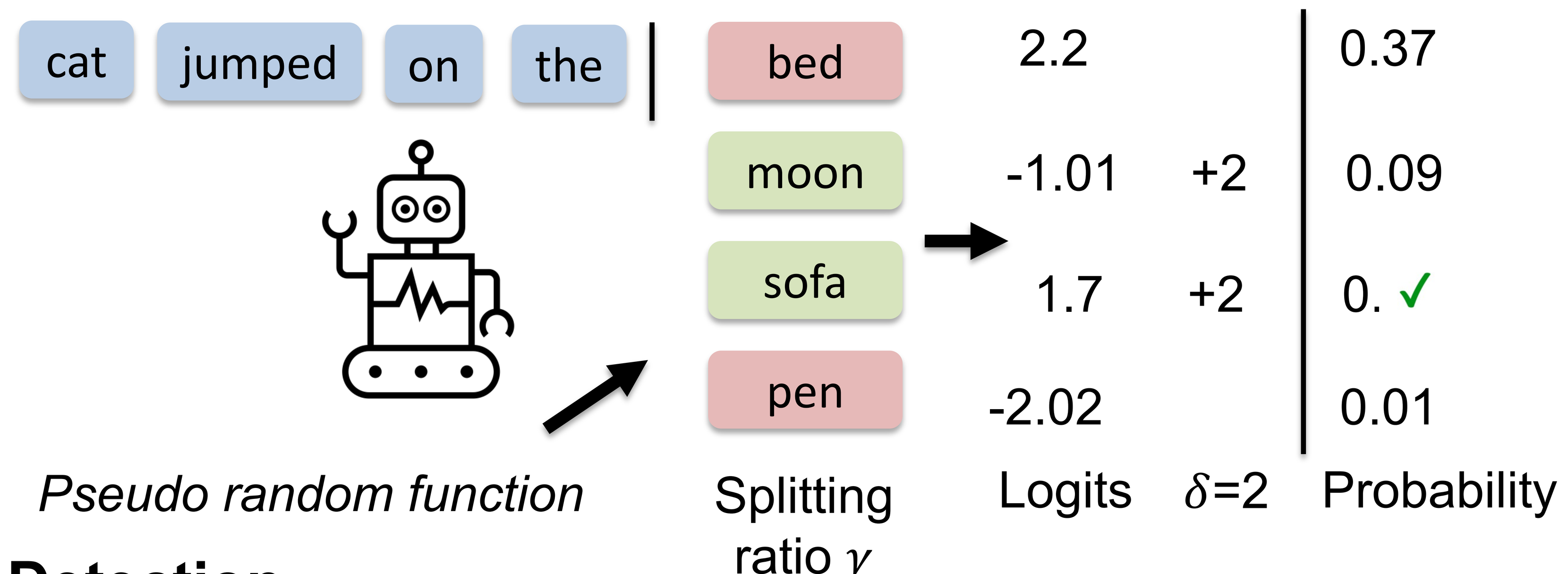
- Academic dishonesty
- Spam content
- Misleading content
- Training degeneration

Prior Method

Distribution shift-based methods – KGW

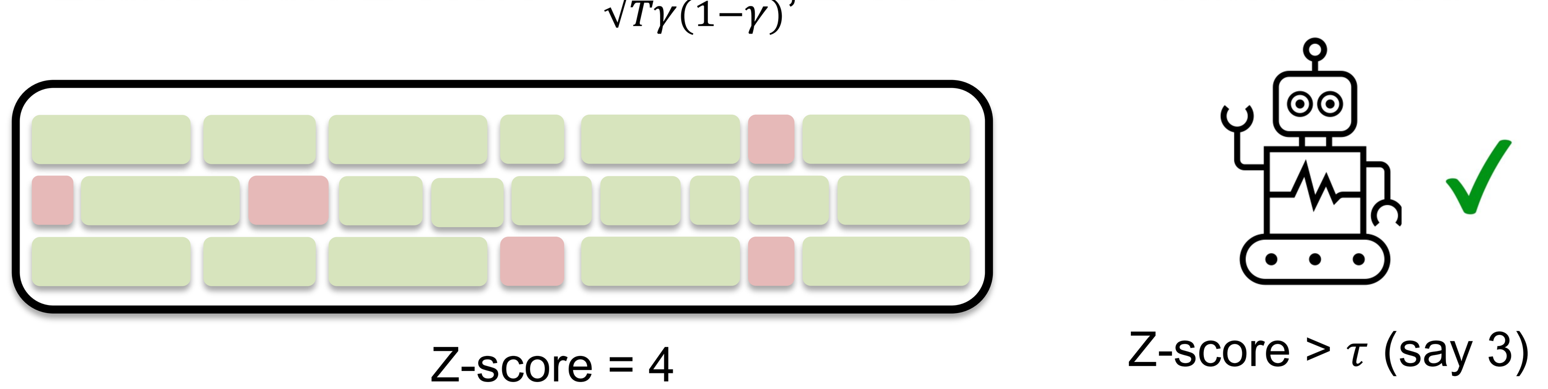
Generation

- Randomly split vocabulary into red-green list with splitting ratio γ
- Bias towards green list words by adding watermark logit $\delta > 0$



Detection

- Count number of green tokens, $|s|_G$, in test sample of length T
- Estimate the Z-score $= \frac{|s|_G - T\gamma}{\sqrt{T\gamma(1-\gamma)}}$; Z-score $> \tau \Rightarrow$ watermarked

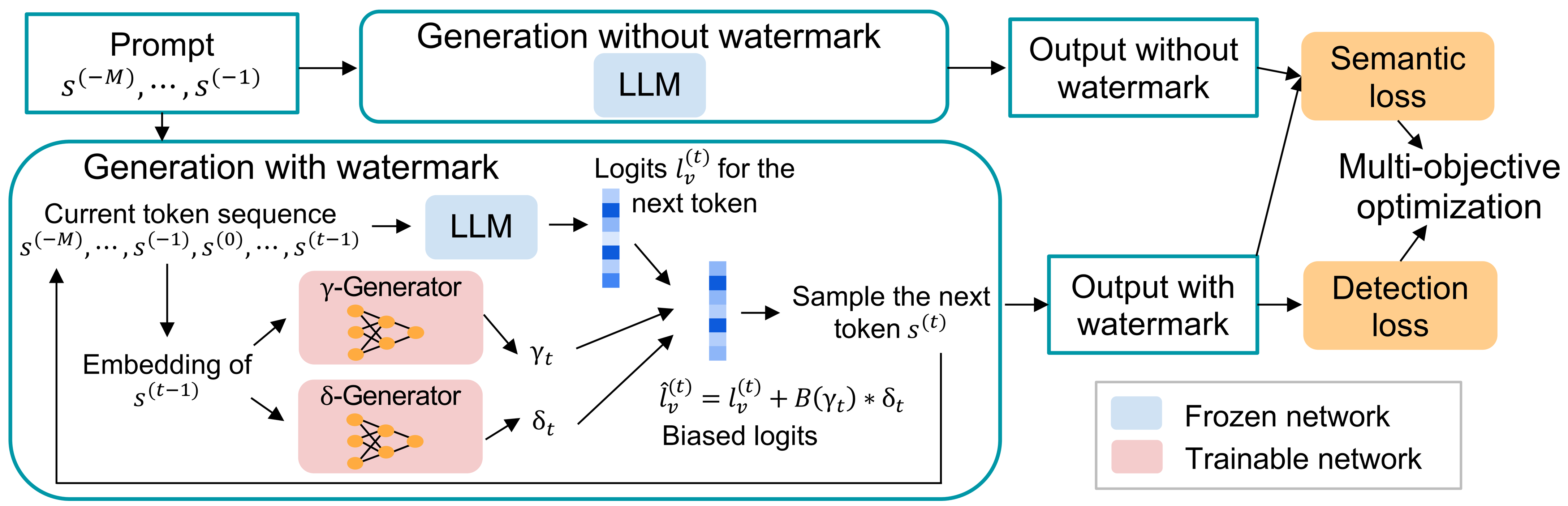


Limitations

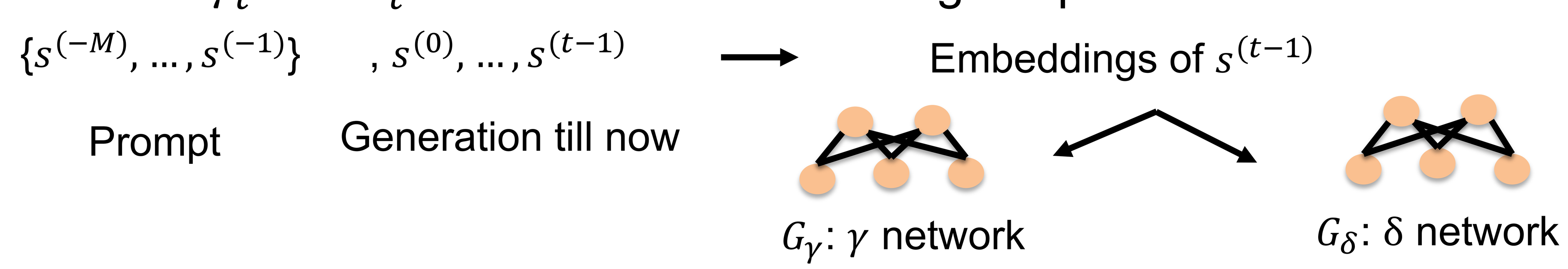
- Cannot simultaneously optimize semantics and detectability
- Lack adaptive mechanism to adjust γ and δ appropriately
  - Sun rises in the __ -> 'east'
  - High δ and low γ might affect semantics

Proposed Method

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



Determine γ_t and δ_t based on the embeddings of previous token



Split the vocabulary (V) into red-green list

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

Bias green list tokens

- Given logits $l_v^{(t)}$, modified logits for token v are: $\hat{l}_v^{(t)} = l_v^{(t)} + y_v^{(t)} * \delta_t$

Training objectives

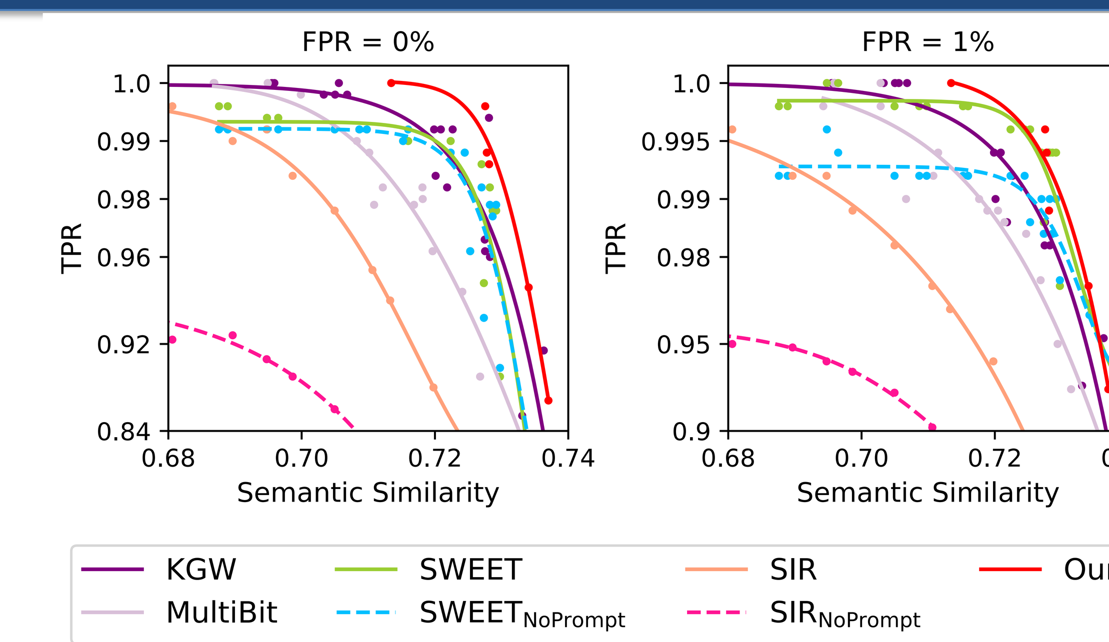
- Detection loss
  - Modified Z-score $= \frac{|s|_G - \sum_{t=1}^{T} \gamma_t}{\sqrt{\sum_{t=1}^{T} \gamma_t(1-\gamma_t)}}$ to account for varying γ_t
  - Improve detectability by maximizing this objective
  - $|s|_G$, count of green tokens, is non-differentiable w.r.t γ_t and δ_t
  - Propose differentiable 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, $L_D = -\hat{z}$
- 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, $L_S = -\cos_{sim}(f_\theta(s),\ f_\theta(s_w))$

Multi-objective optimization

- Optimizing for two competing loss functions L_D and L_S

  $\min_{G_\gamma, G_\delta} L_D(G_\gamma, G_\delta)$ and $\min_{G_\gamma, G_\delta} L_S(G_\gamma, G_\delta)$
- Estimate pareto optimal solutions using multiple-gradient descent algorithm (MGDA)

Experimental Results



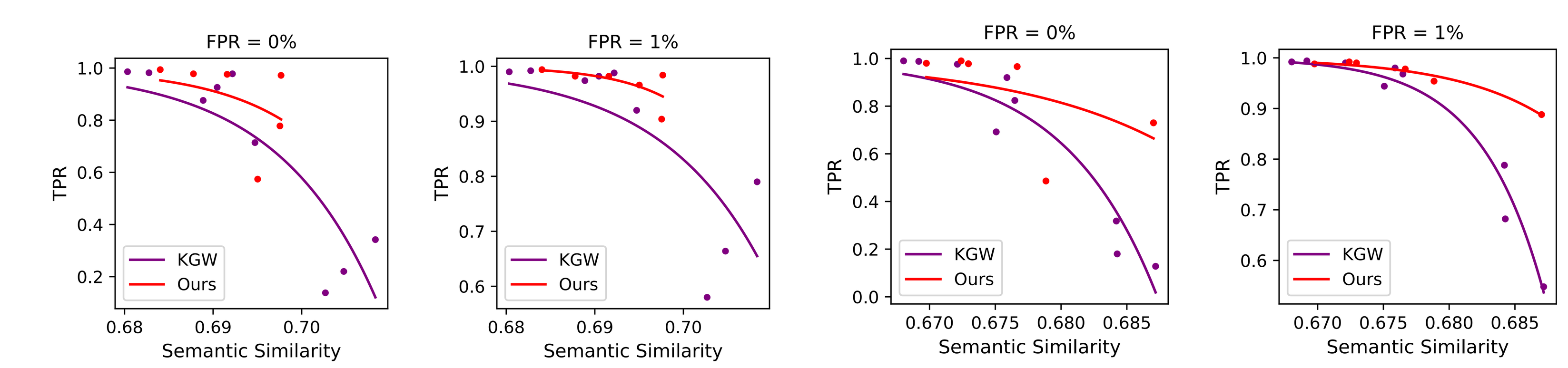
1. Trade-off curves for our method and other baselines applied to OPT-1.3B.

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**

2. Comparison of our method with indistinguishable method - EXP-edit

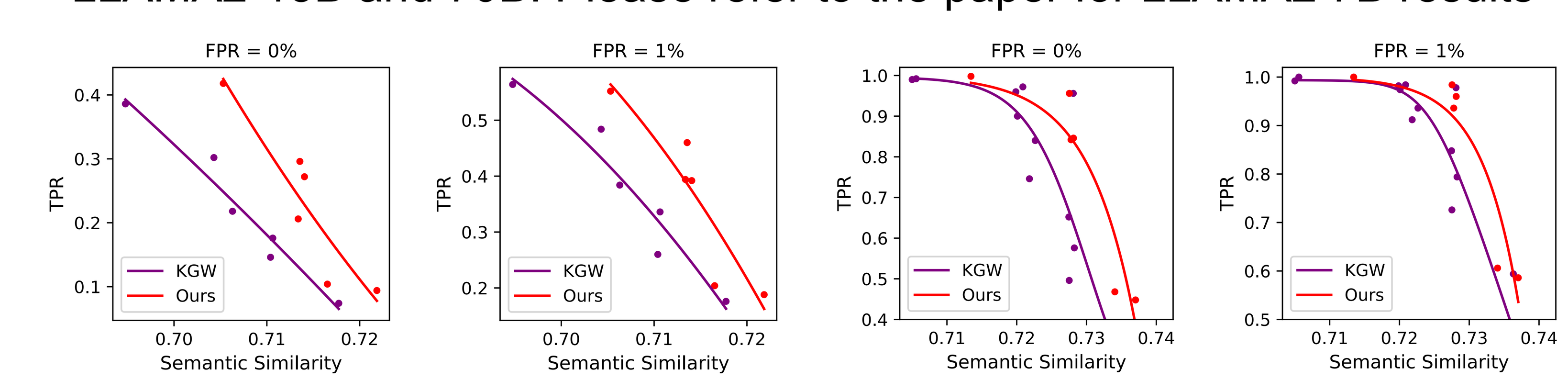
Method	Generation (s)	Detection (s)
No Watermark	3.220	-
KGW	3.827	0.067
SWEET	4.030	0.127
EXP-edit	24.693	155.045
SIR	8.420	0.337
MultiBit	6.500	0.610
Ours	3.946	0.166

3. Generation and detection time.



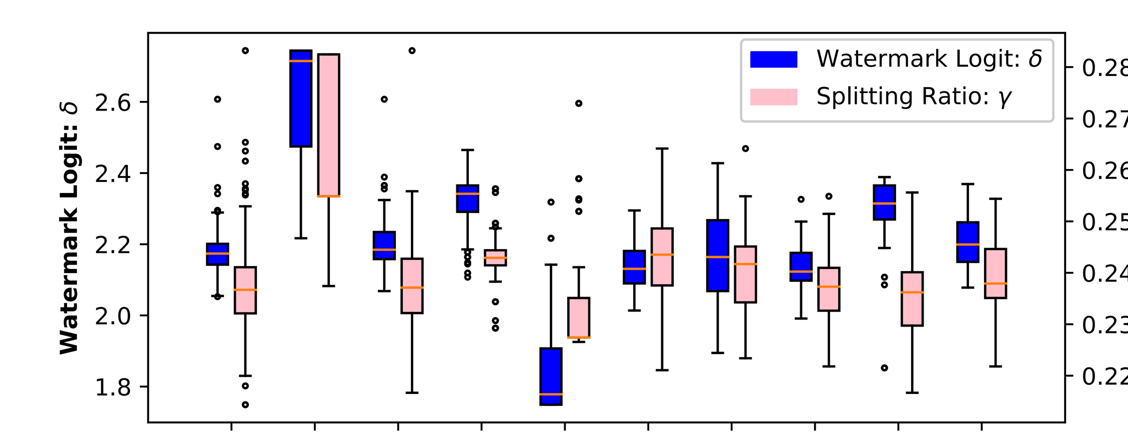
a. LLAMA2-13B b. LLAMA2-70B

4. Performance of our model (trained on OPT-1.3B) and KGW when applied to LLAMA2-13B and 70B. Please refer to the paper for LLAMA2 7B results



a. Dipper paraphrase attack b. Copy-paste-3 attack

5. 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.



6. Distribution of δ (left y-axis) and γ (right y-axis) across different part-of-speech categories of the preceding token.