Segmenting Watermarked Texts From Language Models

Xingchi Li[1], Guanxun Li[2], Xianyang Zhang[1]

[1]Texas A&M University, [2]Beijing Normal University at Zhuhai

Basic Setups

- Prompts and tokens generated before time t:

$$y_{-n_0:t-1} = \underbrace{y_{-n_0} \cdots y_0}_{\text{prompt}} \ \underbrace{y_1 \cdots y_{t-1}}_{\text{tokens generated}} \in \mathcal{V}^{t+n_0},$$

  where $\mathcal{V}$ is the vocabulary.

- Next token distribution: an LLM maps $y_{-n_0:t-1}$ to a distribution over the vocabulary

$$p_t(\cdot) := p(\cdot | y_{-n_0:t-1})$$

  for generating the next token y_t.

Watermarking: the basic idea

A watermark text generation algorithm recursively generates a string $y_{1:n}$ by

$$y_t = \Gamma(\xi_t, p_t(\cdot)), \quad 1 \leq t \leq n,$$
$$\xi_t \sim q, \quad 1 \leq t \leq n.$$

- ξ_t is the watermark key that will be provided to the detector.
- Γ is a decoder function.

Distortion-free

The watermarking scheme preserves the original text distribution:

$$P(\Gamma(\xi_t, p_t) = y) = p_t(y).$$

A Hypothesis Testing Problem

- Given the published text $\widetilde{y}_{1:m}$ and the watermark key sequence $\xi_{1:n}$, the detector tests

  $\mathbb{H}_0 : \widetilde{y}_{1:m}$ is non-watermarked vs $\mathbb{H}_a : \widetilde{y}_{1:m}$ is watermarked.

- Under $\mathbb{H}_0$, $\widetilde{y}_{1:m}$ and $\xi_{1:n}$ are independent.

- Under $\mathbb{H}_a$, $\widetilde{y}_{1:m}$ and $\xi_{1:n}$ are statistically dependent.

A Randomization Test

- Let $\phi(\xi_{1:n}, \widetilde{y}_{1:m})$ be a test statistic, which measures the dependence between $\xi_{1:n}$ and $\widetilde{y}_{1:m}$. Examples include
  1. Pearson correlation
  2. Rank correlation
  3. Levenshtein distance
  4. Edit distance

- Generate $\xi_t^{(b)} \sim q$ independently over $1 \leq t \leq n$ and $1 \leq b \leq B$.

- The p-value is given by

$$p_B = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\phi(\xi_{1:n}, \widetilde{y}_{1:m}) \leq \phi(\xi_{1:n}^{(b)}, \widetilde{y}_{1:m})\}}{B+1}.$$

Type I and Type II Error Control

Theorem

(i) *Under the null,* $P(p_B \leq \alpha) = \lfloor (B+1)\alpha \rfloor / (B+1) \leq \alpha$;

(ii) *Suppose the following three conditions hold:*

(a) $\max\{\mathrm{Var}(\phi(\xi_{1:n}, \widetilde{y}_{1:m})|\mathcal{F}_m), \mathrm{Var}(\phi(\xi'_{1:n}, \widetilde{y}_{1:m})|\mathcal{F}_m)\} \leq C/n$;
(b) $\mathbb{E}[\phi(\xi'_{1:n}, \widetilde{y}_{1:m})|\mathcal{F}_m] = O(n^{-1/2})$;
(c) $\lim_{n\to\infty} \sqrt{n}\mathbb{E}[\phi(\xi_{1:n}, \widetilde{y}_{1:m})|\mathcal{F}_m] = \infty$.

$\mathcal{F}_m = [y_{-n_0:0}, \widetilde{y}_{1:m}]$ *and* $\xi'_{1:n}$ *(independent of* $\widetilde{y}_{1:m}$*) is generated in the same way as* $\xi_{1:n}$*. For any* $\epsilon > 0$*, when* $B > 2/\epsilon - 1$*,*

$$P(p_B \leq \alpha|\mathcal{F}_m) \geq 1 - D\exp(-2B\epsilon^2) + o(1), \tag{1}$$

as $n \to +\infty$*, where* $D > 0$.

Higher Entropy, Easier Detection

Corollary

For exponential minimum and inverse transform sampling, (1) holds when

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - p(y_i | y_{-n_0:i-1})\right) \to \infty.$$

Fact: Texts from more advanced LLMs tend to have lower entropy, making detection harder.

Segmenting Watermarked Text

Question: Can we identify the sub-strings from the modified text $\widetilde{y}_{1:m}$ that are machine-generated?

Assumption: The text published by the user has the structure:

$$\underbrace{\widetilde{y}_1\widetilde{y}_2\cdots\widetilde{y}_{\tau_1}}_{\text{non-watermarked}}\ \underbrace{\widetilde{y}_{\tau_1+1}\cdots\widetilde{y}_{\tau_2}}_{\text{watermarked}}\ \underbrace{\widetilde{y}_{\tau_2+1}\cdots\widetilde{y}_{\tau_3}}_{\text{non-watermarked}}\ \underbrace{\widetilde{y}_{\tau_3+1}\cdots\widetilde{y}_{\tau_4}}_{\text{watermarked}}\cdots$$

Goal: Separate the text into watermarked and non-watermarked sub-strings accurately

Insight: Turn it into a change-point problem

Moving Windows

- Define a sequence of moving windows:

$$\mathcal{I}_i = [(i - B/2) \vee 1, (i + B/2) \wedge m].$$

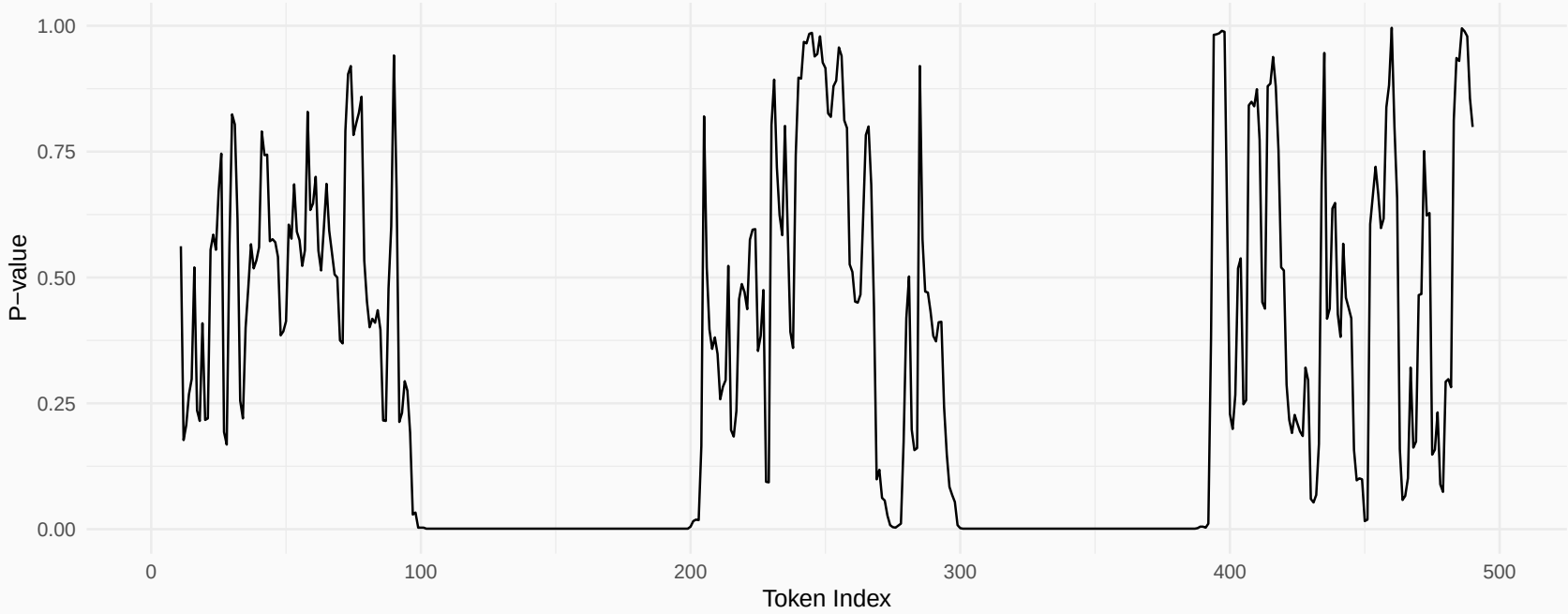
- Compute the randomization-based p-value:

$$p_i = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbf{1}\{\phi(\xi_{1:n}, \widetilde{y}_{\mathcal{I}_i}) \leq \phi(\xi_{1:n}^{(b)}, \widetilde{y}_{\mathcal{I}_i})\}\right).$$

A Newspaper Example

Local electronics retailers complain the deal gives Best Buy an unfair advantage. Citizen groups such as Recycle, Act Orange, and Green in the Charlottesville Region have said they oppose the deal because it does little to promote capital spending to build the city's economy. Joy Phillips, a local businessman who has fought so loudly that he has earned a column in the local newspaper all about him, summed up his take on the deal in a recent letter to the editor: "No price for local businesses, almost no price for a national retailer that currently exists in Charlottesville, and only a few new employees." Trash Tax After All: Remember when the Charlottesville City Council said it was going to tax trash? don't worry; the recyclingwatch.org intelligentsia won't be able to fight this one. After all, recycling is a hobby (for some), not an economic driver. Today, however, Councilmember Rob Schilling argued for rolling back the city's tax on trash pickup because the cost of collection, particularly container policies, has soared over the past eight years. But his fellow Council members voted down his motion to revisit the issue, and the two new Council members voted with the proposal's critics ..... Good news: There were two No votes from ..... discusses Henry Waxman's proposal to cut off taxpayer funding to the National Endowment for the Arts And that is what it is .....

A Newspaper Example



A Single Change-point

- Test the distributional change in the p-value sequence $p_1, \ldots, p_m$.

- Test statistic:

$$T_m = \max_{1 \le \tau < m} S_{1:m}(\tau)$$

  with

$$S_{1:m}(\tau) := \sup_{t \in [0,1]} \frac{\tau(m-\tau)}{m^{3/2}} |F_{1:\tau}(t) - F_{\tau+1:m}(t)|$$

- Change-point estimate:

$$\hat{\tau} = \arg\max_{1 \le \tau < m} S_{1:m}(\tau)$$

Block Bootstrap

- The p-value sequence is B-dependent: p_i and p_j are independent if and only if $|i - j| > B$.

- We use block bootstrap to resample a sequence of p-values $p_1^*, \ldots, p_m^*$ and recompute the test statistic.

- Define the block bootstrap-based p-value

$$\widetilde{p} = \frac{1}{B' + 1} \left(1 + \sum_{b=1}^{B'} \mathbf{1} \left\{ T_m \leq T_m^{*,(b)} \right\} \right).$$

  We claim that there is a statistically significant change point if $\widetilde{p} \leq \alpha$.

Consistency

Theorem

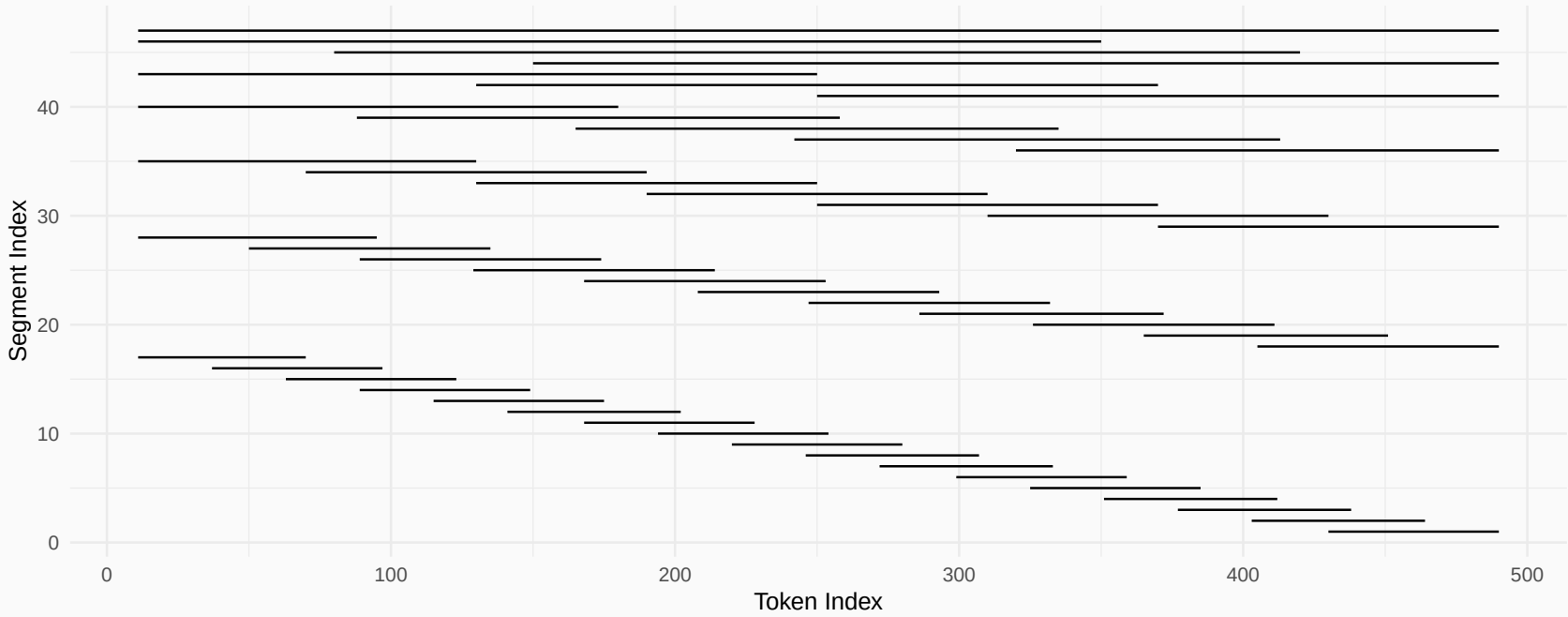
When there is a single change point located at τ^*,

$$|\hat{\tau} - \tau^*| = O_p\left(\frac{\sqrt{mB\log(m/B)}}{D(F_0, \mathbb{E}[F_{\tau^*+1:m}(t)])}\right),$$

where $D(F, G)$ denotes the Kolmogorov–Smirnov distance between F and G.

Remark. The p-values from the watermarked segments could follow different distributions.

Seeded Intervals With Varying Lengths



Algorithm 1

Require: p-values $\{p_i\}_{i=1}^m$, threshold ζ, seeded intervals $\mathcal{I}$.

for $i \leftarrow 1, \ldots, |\mathcal{I}|$ **do**

For the i-th interval $I_i = (r_i, s_i]$, compute the change point estimate $\hat{\tau}_i = \arg\max_{r_i < \tau \leq s_i} S_{r_i+1:s_i}(\tau)$, where

$$S_{r_i+1:s_i}(\tau) := \sup_{t \in [0,1]} \frac{(\tau - r_i)(s_i - \tau)}{(s_i - r_i)^{3/2}} |F_{r_i+1:\tau}(t) - F_{\tau+1:s_i}(t)|.$$

Obtain $\widetilde{p}_i$ through block bootstrap.

end for

Potential and final change points $\mathcal{O} = \{i : \widetilde{p}_i < \zeta\}$, $\mathcal{S} = \emptyset$.

while $\mathcal{O} \neq \emptyset$ **do**

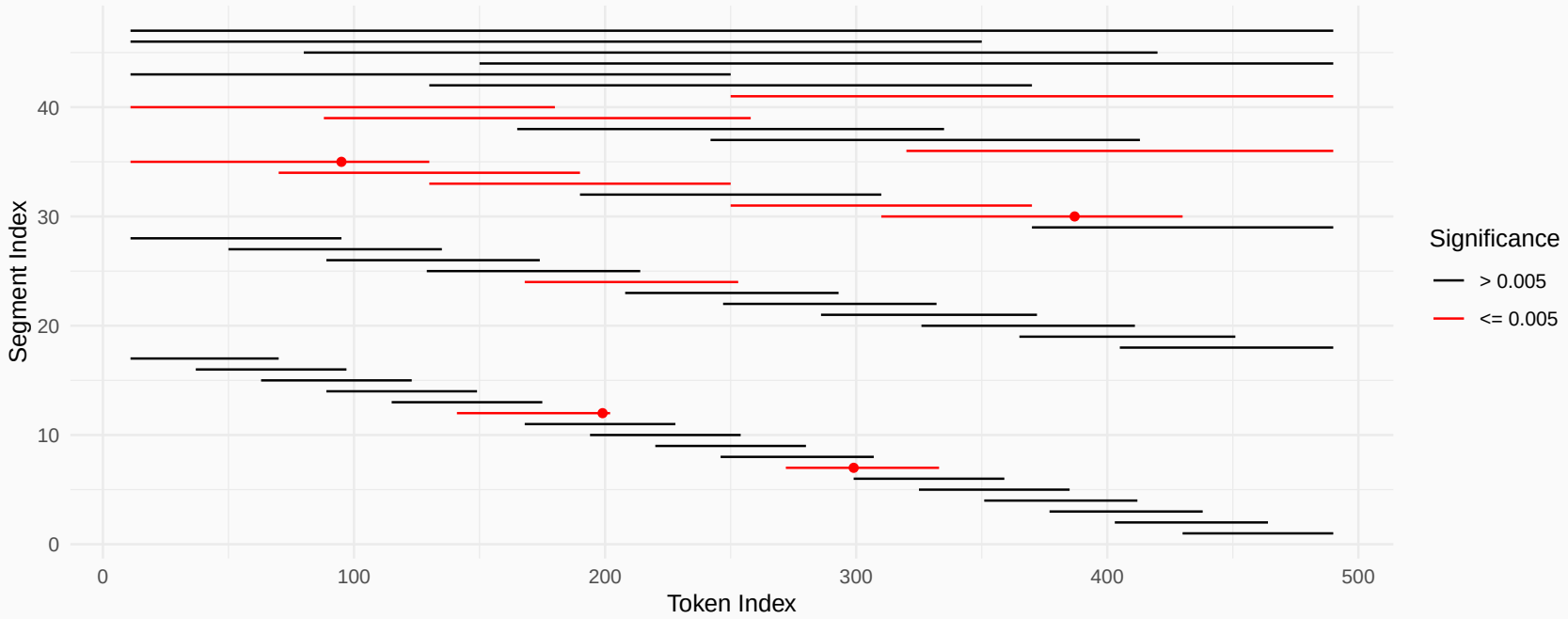
Select $i = \arg\min_{i=1,\ldots,|\mathcal{O}|}\{|I_i|\} = \arg\min_{i=1,\ldots,|\mathcal{O}|}\{s_i - r_i\}$.

$\mathcal{S} \leftarrow \mathcal{S} \cup \{\hat{\tau}_i\}$; $\quad \mathcal{O} \leftarrow \{j \leq |\mathcal{O}| : \hat{\tau}_i \notin I_j\}$.

end while

return $\mathcal{S}$.

A Newspaper Example



Experimental Settings

- Setting 1: Generate 500 tokens with a watermark.
- Setting 2: Generate 250 tokens with a watermark and append with 250 tokens without watermark.
- Setting 3: Generate 500 tokens with a watermark and substitute the token with indices ranging from 201 to 300 with non-watermarked text.
- Setting 4: Generate 400 tokens with a watermark, substitute the token with indices ranging from 101 to 200 with non-watermarked text, and insert 100 tokens without watermark at the index 300.

False Positives

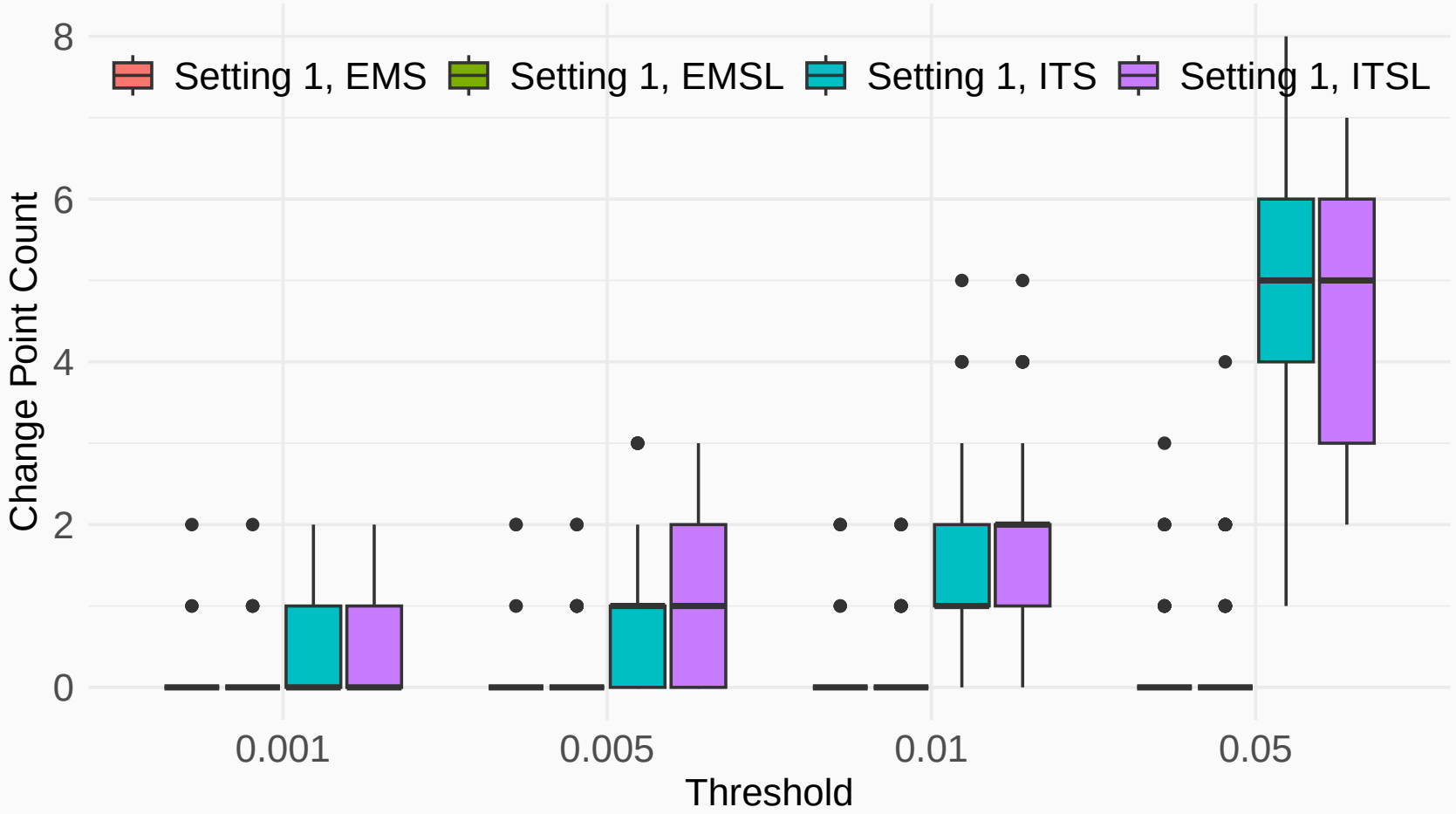


Figure 1: Boxplots of the number of false positives.

Estimation Accuracy

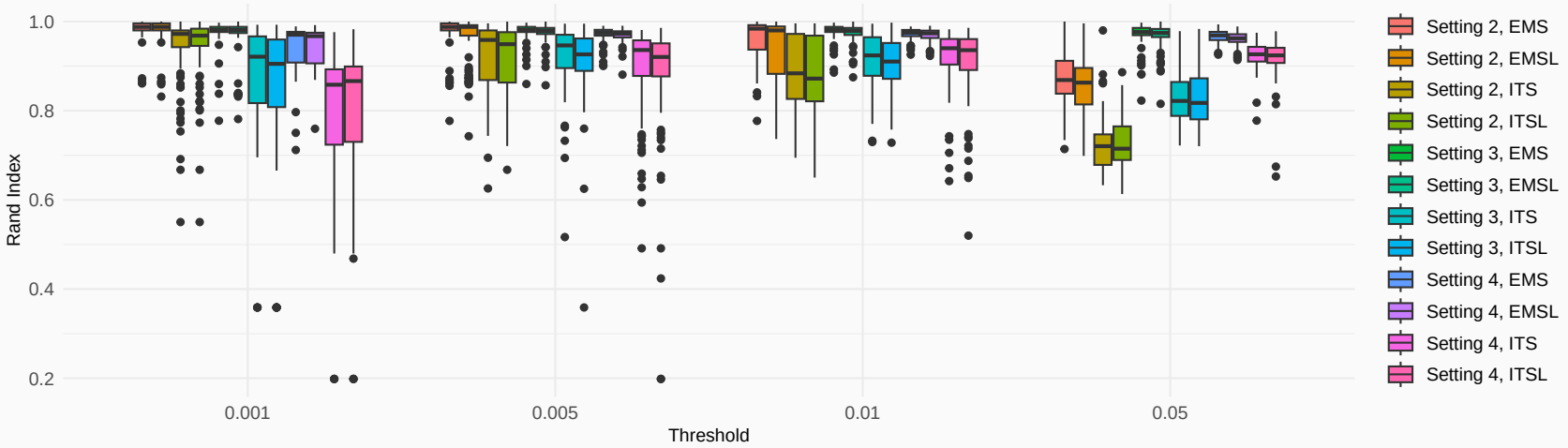


Figure 2: Boxplots of the Rand index comparing the identified and true clusters.

Discussions

- Optimal watermarking and detection schemes
- Watermarked texts from multiple LLMs
- Other types of attacks, e.g., rephrase using a different LLM

Thank you!