Attacking Watermarks for Large Language Models

UC SANTA BARBARA
Early Research Scholars Program

Emerson Yu | Zeel Patel | Siddhi Mundhra | Brian Sen
Advisor: Prof. Prabhanjan Ananth

Motivation

- Watermarking is crucial for discerning AI-generated vs. human text.
- Existing watermarking schemes are vulnerable to simple attacks.
- Evaluating robustness empirically guides improvements in detection.

Goals

- Test robustness of 2 watermarks against attacks.
- Methodically expose their green and red lists.
- Analyze effect of output length on detection accuracy.

Methodology

- On LLM OPT 1.3B, used custom pipelines to prompt and detect model with Soft/Unigram/no watermark.
- Used Google Colab's T4 GPU and 1 local GPU.

Background

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

Experimental Results

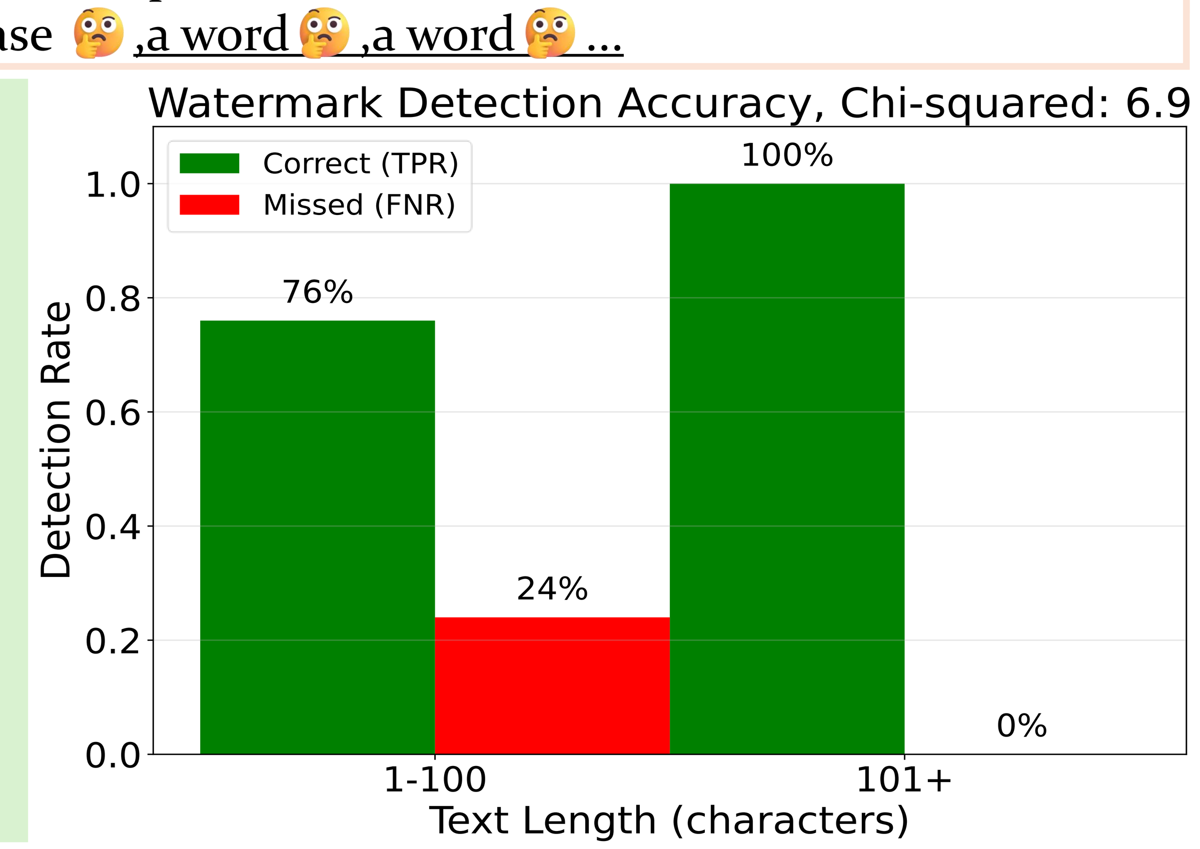
Attack Examples: Generated 160 attack inputs using ChatGPT, including: space substitutions, linguistic attacks, & encodings (e.g. emojis, 1st vs. 3rd person, & Caesar cipher respectively).

- Emoji attack caused false negatives (40% for Soft, 60% for Unigram) by disrupting Soft detector's regeneration of lists and directly erasing green tokens in Unigram output.

Prompt: Distant 🫣 thunder 🫣 murmurs 🫣 warnings 🫣 over 🫣 landscapes 🫣 shrouded 🫣 in 🫣 eerie 🫣 silence
Output *(Soft)*: 🫣 wind 🫣 mood 🫣 ,aestheticallybeautifull 🫣 ...,a phrase 🫣 ,a word 🫣 ,a word 🫣 ...

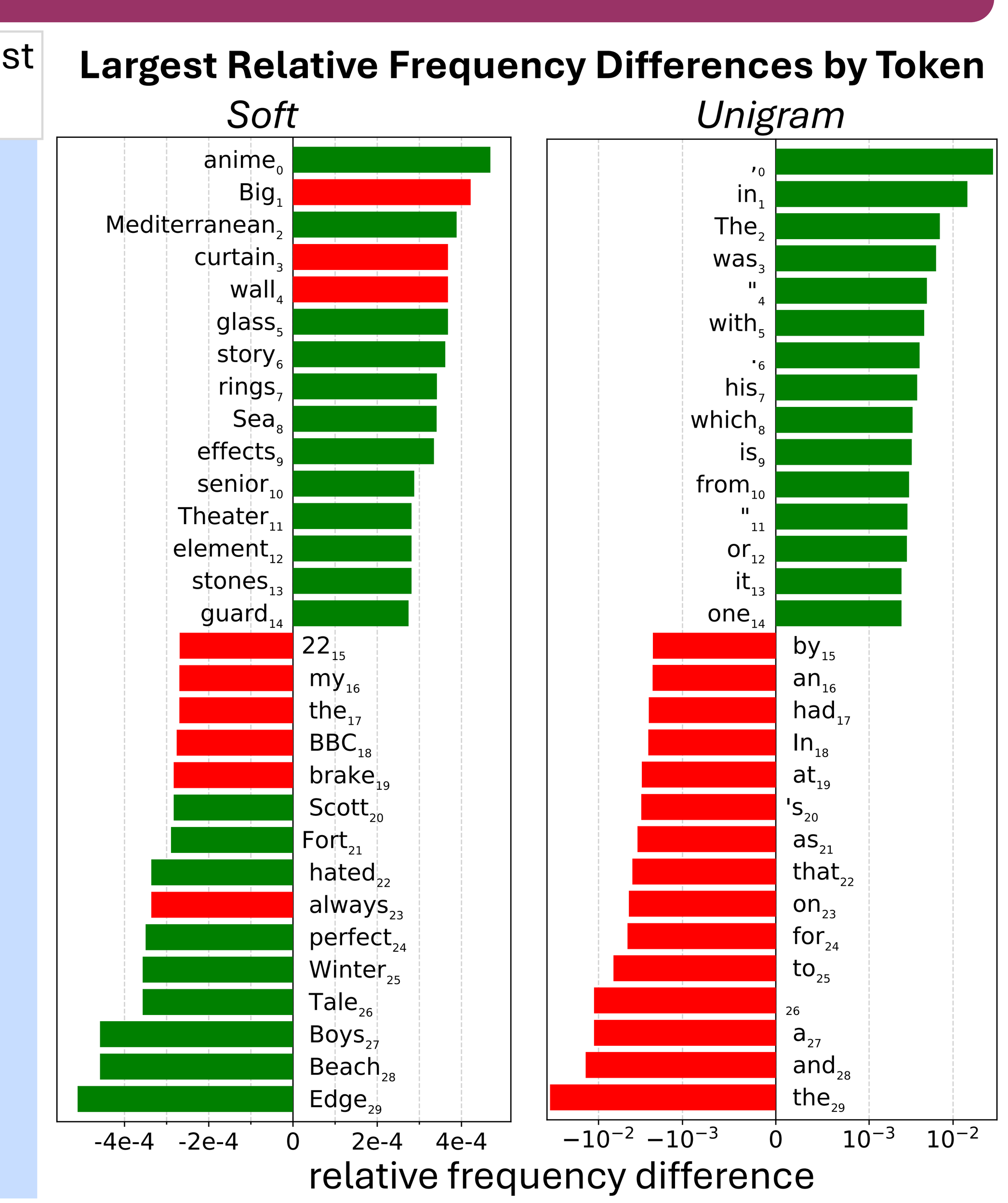
Detection Accuracy vs. Output Length

- Noted most false negatives occurred in shorter watermarked texts.
- Explored relationship between token length and detection accuracy and found longer texts = higher entropy and better detection.
- Chi-squared statistic of 6.9 indicates significant relationship between output length and detection.
- Few false negatives overall; mainly in low character/word count samples.



Frequency Distribution Analysis

- Used 500 OpenGen prompts.
- Relative frequency differences = watermarked-unwatermarked token distributions.
- Tested accuracy of assumption that largest 50 positive/negative differences were green/red tokens, respectively.
- 70% (80/60%) accuracy for Soft vs. 98% (100/96%) accuracy for Unigram.
- Higher accuracy on Unigram due to lists staying constant for all tokens.
- Alternative prompt choices (LQFA, 500x same prompt) led to lower accuracy.



Future Plans

- Invent adaptive attacks (sequencing prompts).
- Evaluate advantages of keyless watermarks.
- Attack watermarks post token distribution analysis.

Citations

- Kirchenbauer, J., et al. "A Watermark for Large Language Models." 1 May 2024.
- Zhao, X., et al. "Provable Robust Watermarking for AI-Generated Text." 13 Oct. 2023.
- Bubble images recreated from "Understanding LLM Decoding Strategies" on Medium.

Acknowledgements

We also thank Prof. Ziad Matni and grad mentor Pranjali Jain for their valuable guidance and support.