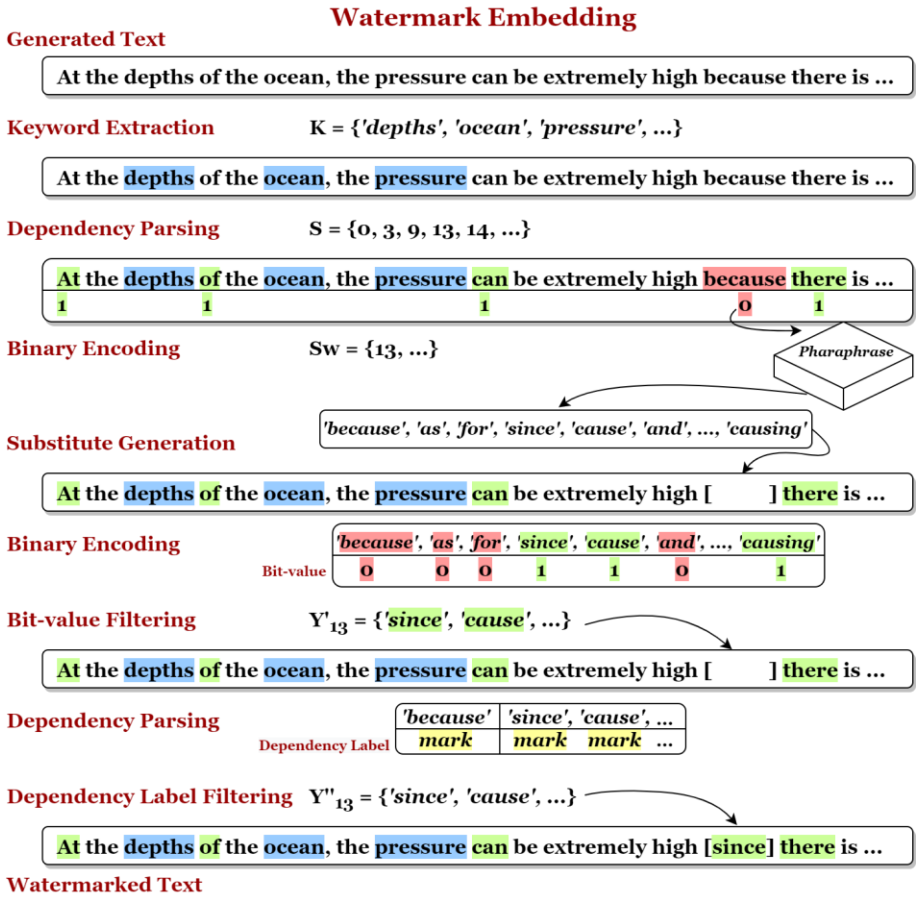
Robust and Semantic-Faithful Post-Hoc Watermarking of Text Generated by Black-Box Language Models

Jifei HAO, Jipeng QIANG, Yi ZHU, Yun LI, Yunhao YUAN, Xiaocheng HU, Xiaoye OUYANG

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40751-w

Problems & Ideas

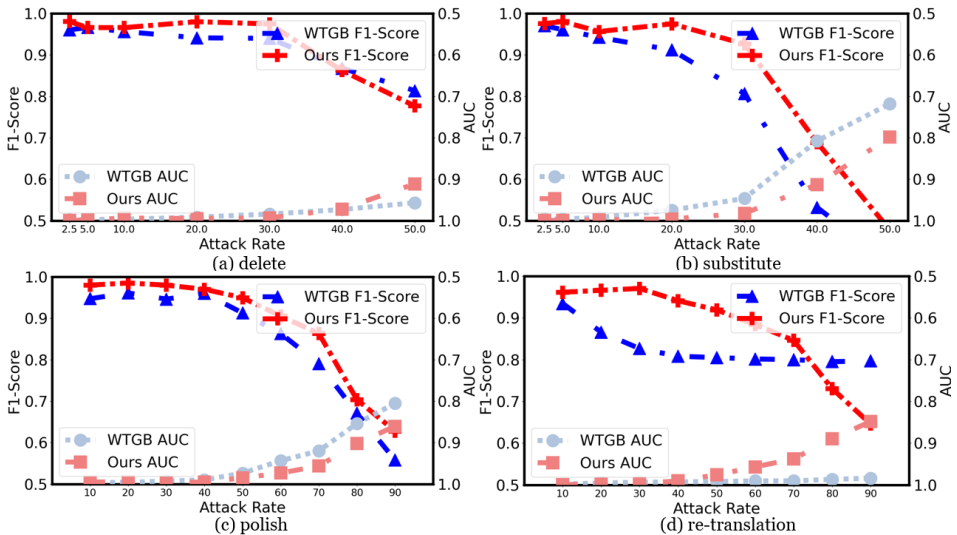
- Problem Overview:
  - **Preserving Semantic Integrity:** Embedding a watermark without altering the original meaning of the text.
  - **Robustness Against Attacks:** Ensuring the watermark is resilient to text modifications, such as word deletion and substitution, which could render traditional watermarking techniques ineffective.
- Core Idea:
  - **Post-hoc Text Watermarking:** Embed watermarks in text generated by black-box language models with minimal modifications, while ensuring high detection accuracy and resilience to various attack scenarios.



The process of the proposing watermark method.

Main Contributions

- Contributions:
  - **Watermark Embedding Framework:** Selects optimal positions using semantic and syntactic features, and encodes watermarks via dependency-based binary encoding, ensuring minimal text modification;
  - **Watermark Detection:** Identifies candidate positions and verifies watermark presence using statistical methods.
  - **Robustness:** Demonstrates resilience to word deletion and substitution attacks, outperforming existing methods in F1 score and detection accuracy.
  - **Experimental Validation:** Evaluated on multiple datasets, showing high watermark strength, detection accuracy, and semantic integrity, even under attack scenarios.



Robustness analysis of the watermark under delete (a), substitute (b), polish (c), and re-translation (d) attacks.

The x-axis represents the attack probability. The y-axis shows the F1-score and AUC for our approach and WTGB. Higher scores indicate better performance.