Signature vs. Substance: Balancing Robustness and Linguistic Quality in LLMs

William Guo[1], Adaku Uchendu[2], and Ana Smith[2]
wguo@imsa.edu, adaku.uchendu@ll.mit.edu, ana.smith@ll.mit.edu
Illinois Math and Science Academy[1] MIT Lincoln Laboratory[2]

Problem Definition

- Large Language Models (LLMs) now produce fluent, human-like text, making it increasingly difficult to identify AI-generated content.
- Text watermarking aims to embed imperceptible patterns into model outputs to enable post-hoc attribution and detection.
- Existing watermarking face a trade-off: greater detectability can degrade linguistic quality. A rigorous evaluation of robustness vs. quality balance is necessary to guide responsible AI deployment.

Research questions:

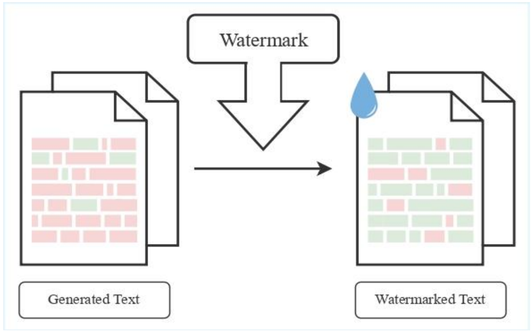
- **RQ1:** How robust are watermarking methods to paraphrasing and backtranslation?
- **RQ2:** How much do they alter linguistic quality?
- **RQ3:** Which linguistic features explain robustness?

Background: LLM Watermarking

- Watermarking: hidden bias in token sampling creates detectable pattern
- Split under “Green list” or “Red list” tokens under secret key
- Detection = statistical test on green-token frequency

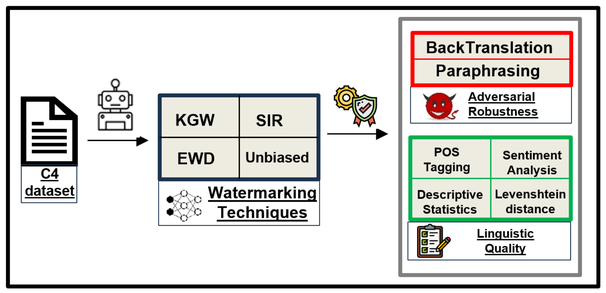
Methods tested:

- **KGW:** adjusting token probabilities
- **EWD:** weight to high-entropy tokens
- **SIR:** semantic embeddings of tokens
- **Unbiased:** shifts the probability distribution



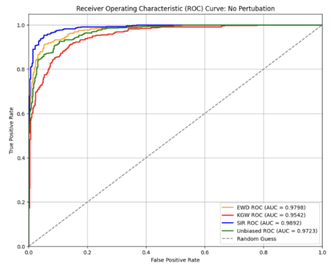
Methodology

10 K passages (50–500 words): C4 corpus → Watermark generation w/OPT-1.3B → LLaMA-3 paraphrase; En↔Fr backtranslation → Detection → Analysis

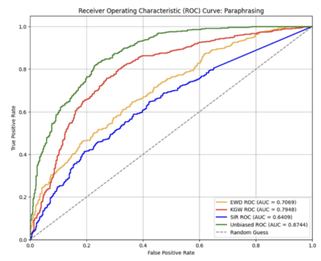


RQ1 Results

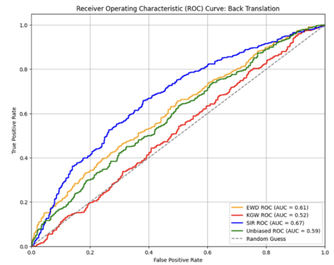
ROC Curves for each watermark under no perturbation, backtranslation, or paraphrase



(a) No perturbation



(b) Paraphrasing

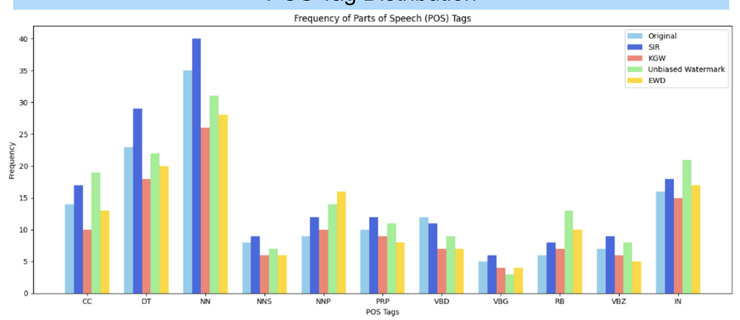


(c) Backtranslation

No attack (AUC ≥ 0.95 (all methods): SIR 0.99 > EWD 0.98 > Unbiased 0.97 > KGW 0.95
Paraphrase (AUC drop −10 to −35 %): Unbiased 0.87 > KGW 0.80 > EWD 0.71 > SIR 0.64
Backtranslation (strongest degradation): SIR 0.67 > EWD 0.61 > Unbiased 0.59 > KGW 0.52

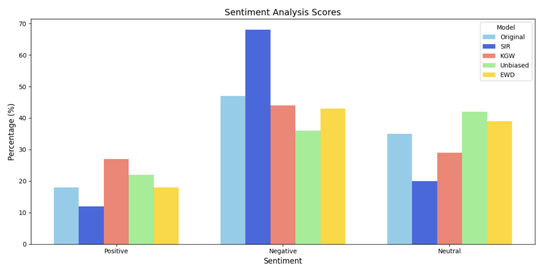
RQ2 Results

POS Tag Distribution



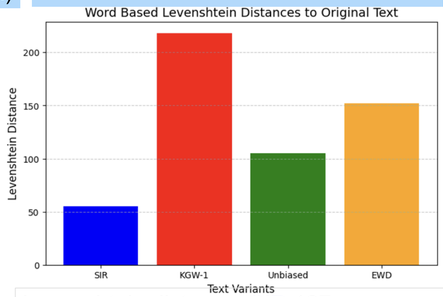
SIR: slight POS shift
Unbiased: minor noun/adverb changes
KGW: most aggressive fewer nouns/determiners
EWD: moderate POS shift, altered nouns/verbs/determiners

Sentiment Analysis % (Positive → Negative → Neutral)



Unbiased best preserves original sentiment balance (neutral/positive ratio). EWD introduces a mild positive bias; KGW and SIR yield more neutral tones.

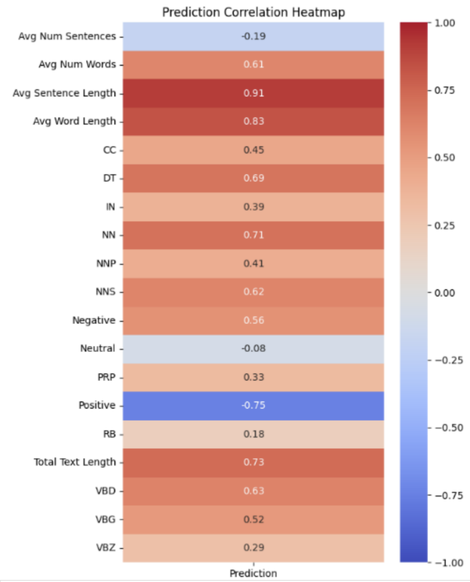
Word Based Levenshtein Distances



Levenshtein distance: SIR 127 < Unbiased 375 < EWD 600 < KGW 823
Sir modifies text minimally, while KGW and EWD introduce more lexical shifts.

RQ3 Results

PCC of linguistic features vs. Predictions



Sentence length r = 0.91, word length r = 0.83, noun ratio r = 0.71. Positive sentiment r = –0.75 lowers robustness. Longer, syntactically complex text better preserves watermark.

Conclusion

- Unbiased watermark is the best balance of detectability + robustness.
- Robustness deteriorates sharply under paraphrasing and translation attacks.
- Longer sentences/words & more nouns correlates to a stronger watermark.
- Positive sentiment → weaker watermark
- Texts with richer syntactic and lexical diversity likely to retain watermark signals

- Future watermarking systems should integrate linguistic awareness and semantic embedding strategies to achieve durability without sacrificing fluency.