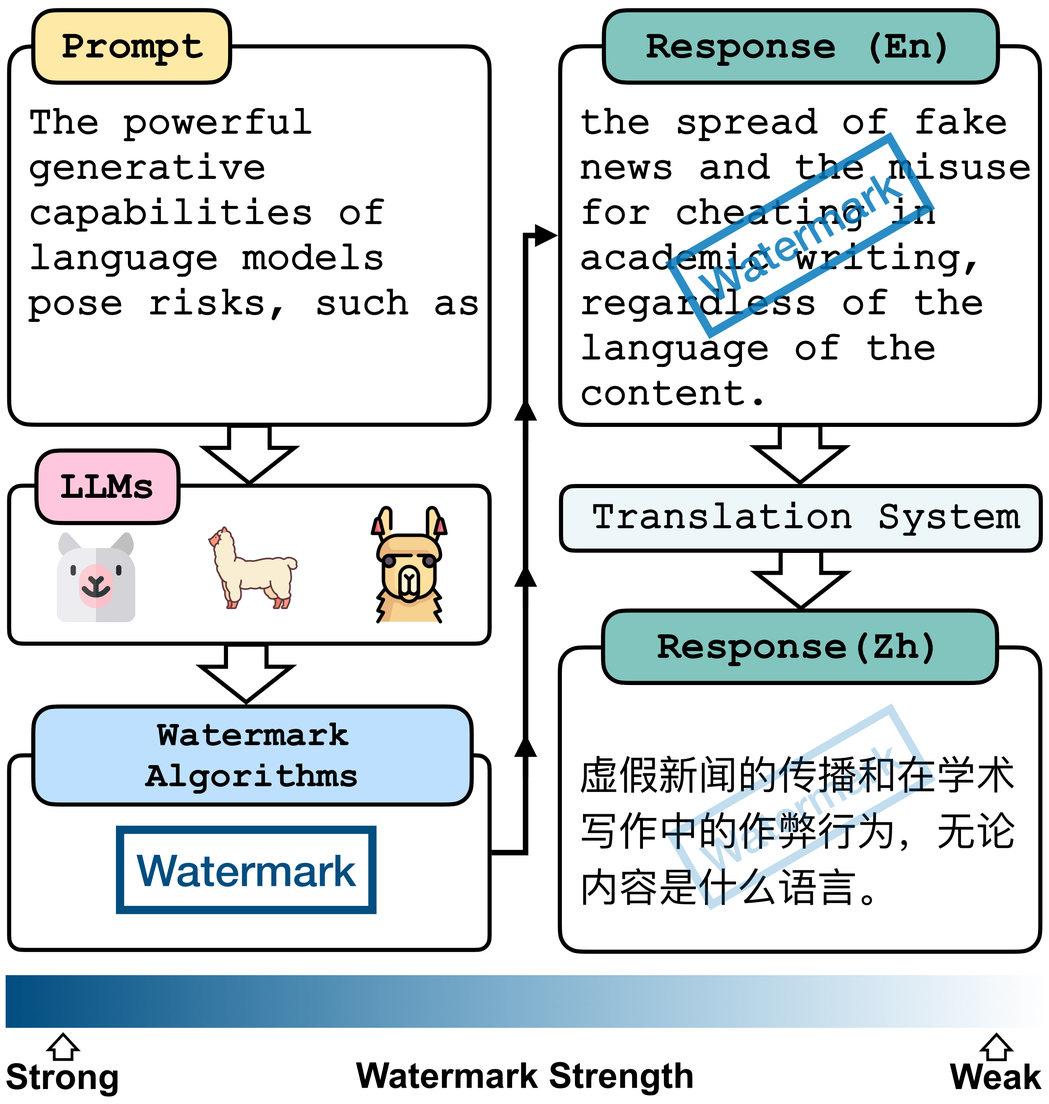
Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models

Zhiwei He#, Binglin Zhou#, Hongkun Hao, Aiwei Liu,
Xing Wang*, Zhaopeng Tu, Zhuosheng Zhang, Rui Wang*

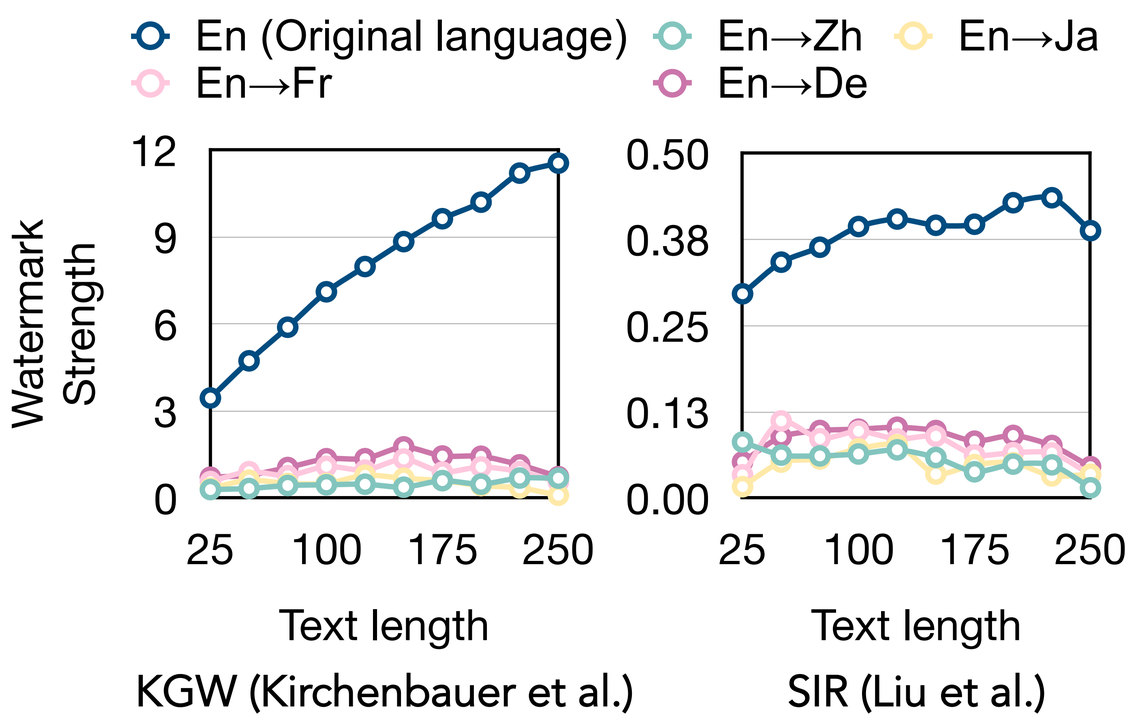
Github

Background & Motivation

- ✔ Text watermaking has been developed to mitigate the misuse of LLMs.
- ✔ It tags and identifies LLM-generated content.
- ? *Can watermarks survive translation* if LLM-generated (watermarked) content is translated into other languages?

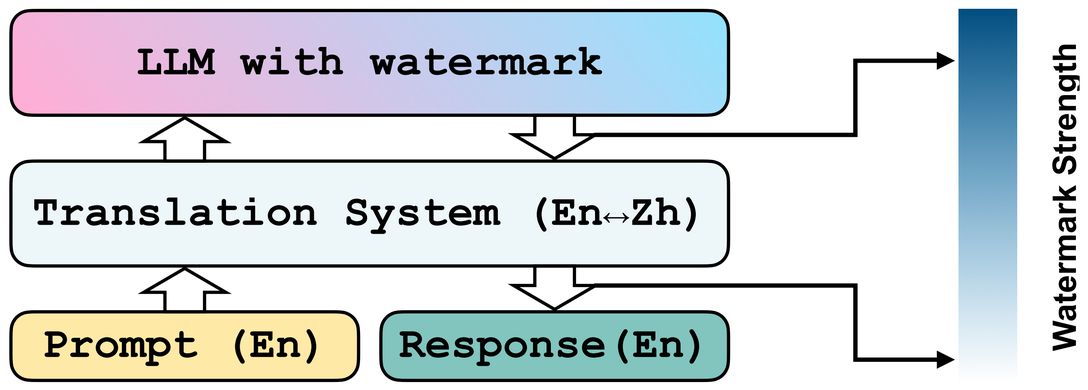


Evaluation: cross-lingual consistency



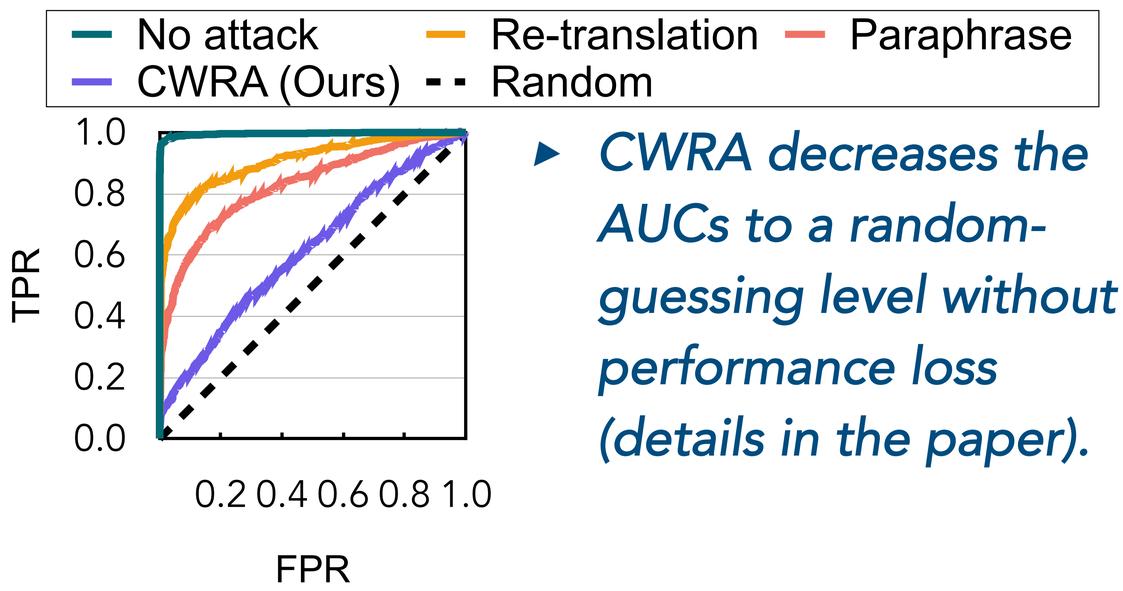
- After translation, the strength of the watermarks drops to almost zero.
  - ▸ *Watermarks cannot survive translation.*

Attack: CWRA method



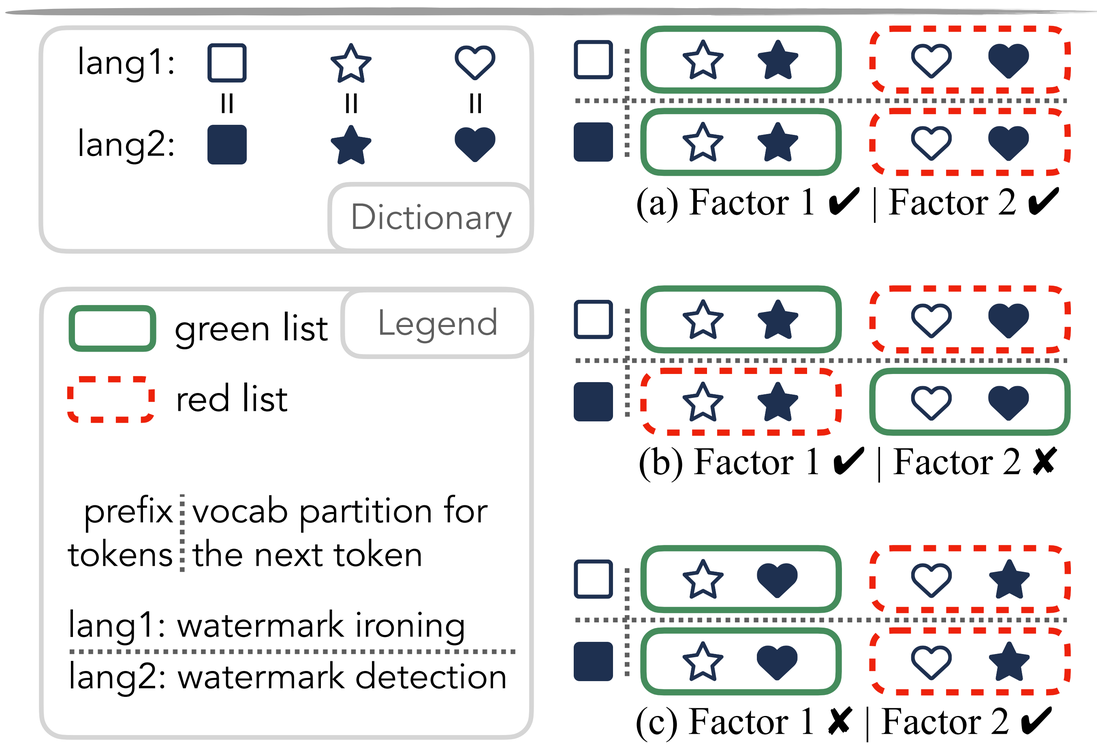
Cross-lingual Watermark Removal Attack (CWRA)

- CWRA wraps the query to the LLM into another language (Zh in the figure).

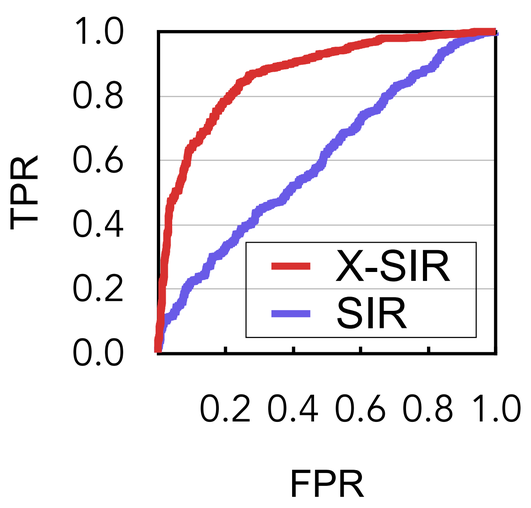


- ▸ *CWRA decreases the AUCs to a random-guessing level without performance loss (details in the paper).*

Defense: X-SIR



- ✔ Factor 1: semantically equivalent tokens must be in the same partition
- ✔ Factor 2: semantically equivalent prefixes must result in the same vocab partitions



- ▸ *By implementing Factor 1 & 2, X-SIR greatly improves watermark detection performance under CWRA.*