XMark: Reliable Multi-Bit Watermarking for LLM-Generated Texts

Scientific Achievement

- Addressed the open question of how to embed reliable, multi-bit watermarks in large language model (LLM)-generated text while preserving text quality and detectability.
- Introduced *XMark*, a method for imperceptible multi-bit watermarking that enables accurate attribution and tracing of generated content.

Significance and Impact

- Enables robust provenance tracking and misuse detection for AI-generated text, supporting safer deployment of LLMs.
- Advances AI security and content authenticity with applications in policy compliance, misinformation mitigation, and industry systems.

Technical Approach

- Encodes binary messages directly into generated text using a structured watermarking scheme that balances detectability and fluency.
- Overcomes challenges of robustness to edits and maintaining text quality, achieving reliable decoding under realistic transformations.

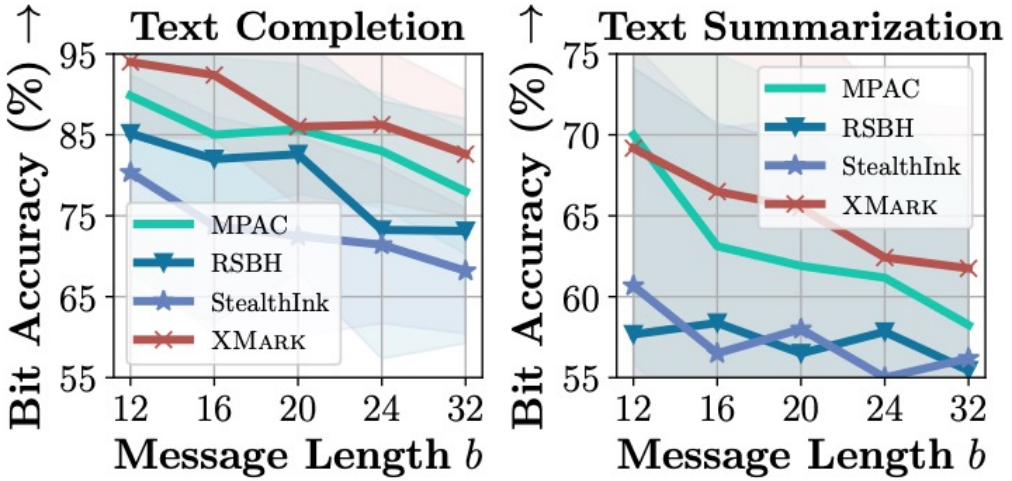

Figure 1: Impact of message length bon Bit Accuracy for MPAC, RSBH, StealthInk, and XMARK on the text completion and text summarization tasks.

PI(s)/Facility Lead(s): K. Kim (PI), O. Kotevska (ORNL's PI)
ASCR Program: AI for Science
ASCR PM: Xujing Davis
Publication(s) for this work: Xu, J., Hu, R., Kotevska, O., Zhang, Z. XMark: Reliable Multi-Bit Watermarking for LLM-Generated Texts. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics.