# Robust Multi-bit Natural Language Watermarking through Invariant Features

> https://arxiv.org/pdf/2305.01904.pdf

**这篇文章用的还是Data Hiding的范式，着重讲在Multi-bit watermarking的情形下如何提升Robustness，这篇文章是在Secret Message的选择上做文章，让Secret Message和原始文本的句法特征有关（Syntatic）**

- Watermarking Process和Watermark Detection Process
  - Sender：获取内嵌了secret message的 $X_{wm}$，需要$X_{wm}$和原始文本$X$尽可能相似，且文本质量不被明显影响
  - Receiver：要从被corrupt的$X_{wm}$中提取出secret message
- Corruptions
  - Step 1：**Word insertion, deletion, substitution** across 2.5% to 5.0% corruption ratios

    Adversarial watermarking transformer: Towards tracing text provenance with data hiding这篇文章里提到过的方法
  - Step 2：为了更好的保证语义变化别太多，它用了一个pretrained sentence transformer all-MiniLM-L6-v2，筛掉了corrupt之后cosine similarity < 0.98
- Framework for Robust Natural Language Watermarking
  - Phase 1：提取Secret Message

    > 这篇文章用的加密范式还是Post-hoc，它提升水印鲁棒性的方法是在Secret Message的选择上做文章，让Secret Message和原始文本的句法特征有关，这样我们可以假设只要不做大幅的破坏，就不会把Secret Message的信息弄丢；
  - Phase 2：把Secret Message嵌入原始文本X得到$X_{wm}$