Geometry-Aware Watermarking of Large Language Models via RoPE

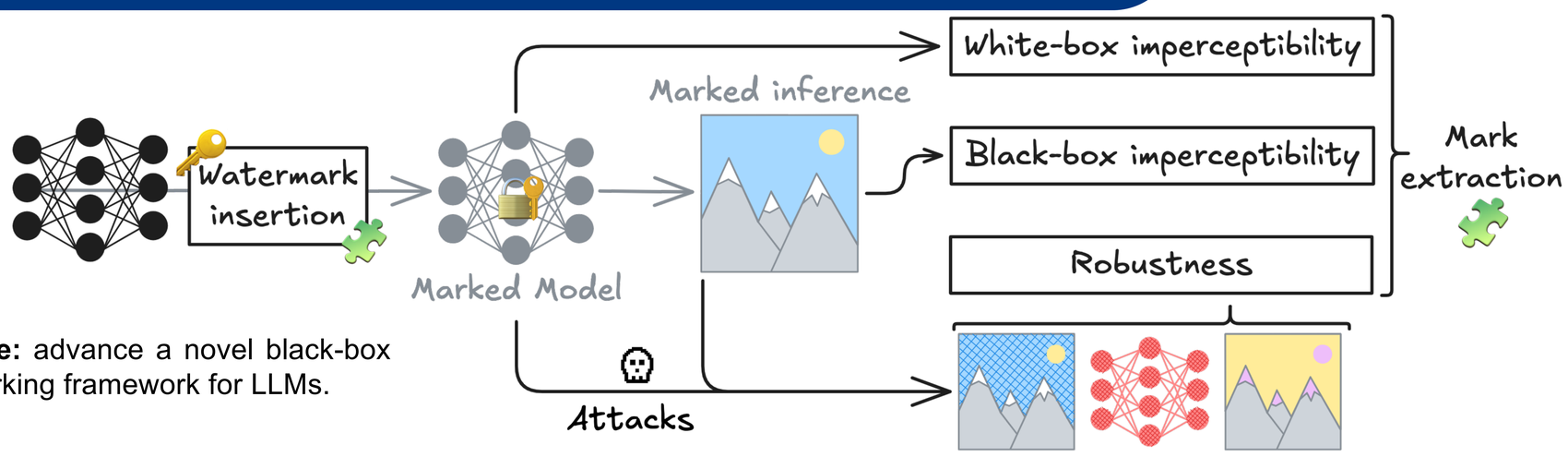
E. Cole, M. Mitrea, H. Chaouchi, SAMOVAR, Telecom SudParis

Research Topic started in Nov. 2025

Context and objective

Protect expensive IP

- **Copyright protection:** LLM development requires massive investments in computing, human intelligence, and proprietary data.
- **The threat:** model usurpation, counterfeiting, or unauthorized redistribution.
- **Objective:** advance a novel black-box watermarking framework for LLMs.



State-of-the-art methods

Passthrough layers

- **Passthrough layers (PTL)** are added to the model modifying its architecture to output random text only when a secret trigger is present.
- **Key characteristics:** does not require access to the model's logits and does not use a detector.
- **Limitations:** the model is modified, and no real key-sensitivity study.

TokenMark

- **TokenMark** leverages the permutation equivariance property of Transformers, the mark being embedded by making the model react differently to permuted inputs.
- **Key characteristics:** structure-based method, so completely model-agnostic.
- **Limitations:** lack of robustness to permutation attacks and no real key-sensitivity study.

Advanced solution

Mathematical Formulation - adding a controlled phase shift

Let $X = [x_1, x_2, \dots, x_i, \dots, x_L] \in \mathcal{R}^{d\times L}$ be the embedded input sequence, and $W_Q, W_k, W_v \in \mathcal{R}^{d\times d}$ be the query, key and value projection matrices, respectively.

$R(i,\theta_k) = \begin{bmatrix} \cos(i\theta_k) & -\sin(i\theta_k) \\ \sin(i\theta_k) & \cos(i\theta_k) \end{bmatrix}$ is a 2D rotation matrix, where $i \in [1, L]$ is the token index and $k \in [0, d/2\,]$ is the k^{th} coordinate pair of x_i hidden dim d.

$R(i) = diag\left[R(i,\theta_0), \dots, R(i,\theta_{d/2})\right]$ is the RoPE matrix, depending on the index.

The queries and keys become with RoPE $\begin{cases} q_i = x_i W_Q, & \tilde{q}_i = x_i W_Q R(i) \\ k_j = x_j W_k, & \tilde{k}_j = x_j W_k R(j) \end{cases}$

The attention is then $\boxed{\tilde{\alpha}_{i,j} = \tilde{q}_i \tilde{k}_j^\intercal = x_i W_Q R(i-j) W_k^\intercal x_j^\intercal}$

Let's divide X into $S = [S_1, \dots, S_N]$, and be $\Delta = [\Delta_1, \dots, \Delta_N]$ a displacement vector, with each $\Delta_m \in \Delta$ the corresponding displacement of all tokens in S_m.

If $(x_{i'}, x_{j'}) \in S_m \times S_n$ then $\begin{cases} i' = i + \sum_{l<m} \Delta_l = i + C_m \\ j' = j + \sum_{l<n} \Delta_l = j + c_n \end{cases}$ where i and j are the indexes without displacement.

Thus, with RoPE and the segments the attention becomes $\boxed{\tilde{\alpha}_{i,j} = \tilde{q}_i \tilde{k}_j^\intercal = x_i W_Q R(i - j + C_m - C_n) W_k^\intercal x_j^\intercal}$ adding a a controlled phase shift depending only on the displacement vector.

Imperceptibility evaluation

	$D_{KL}(P \mid Q) \downarrow$	$D_{JS}(P \mid Q) \downarrow$
(Bl | Wm)	0.013	0.003
(Van | Bl)	0.186	0.040
(Van | Wm)	0.196	0.042

	Vanilla (GPT2)	*PTL*	*Token Mark*	*Ours*
Perplexity ↓	22.88	51.80	30.14	24.02
NLL ↓	3.130	–	–	3.179

KL/JS quantify distribution shift between models (↓ better). Perplexity/NLL measure language modeling quality (↓ better). Van = vanilla model, Bl = baseline watermark, Wm = our method; PTL and TokenMark are state-of-the-art watermarking baselines.

Advanced solution

Watermarking instantiation - RoPE geometric patterns

- **Geometric Triggering:** non-semantic signature injection into the latent space.
- **Detection:** mappings latent states to a signature space, by maximizing **cosine similarity** with a mark S_k only when the RoPE displacement is triggered.

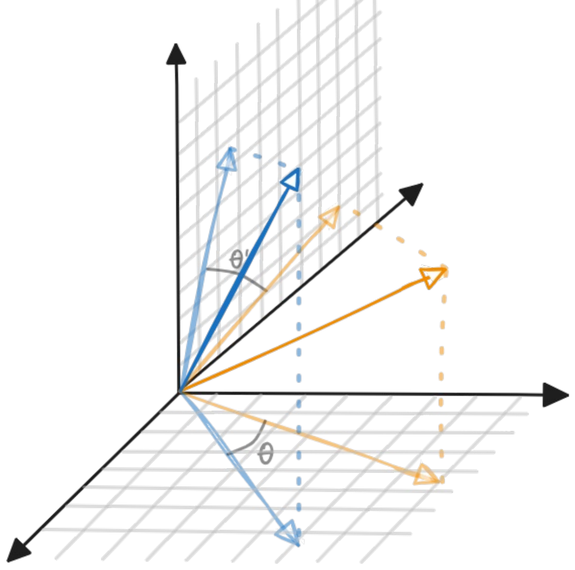
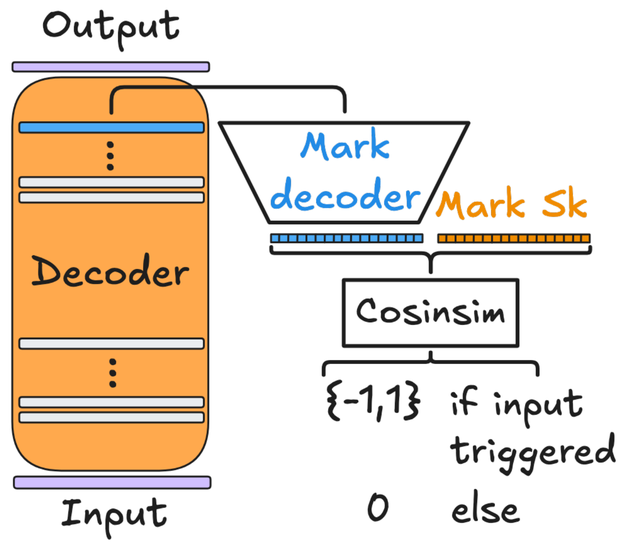


Illustration of Rope being applied to an embedding in a 4d latent space.



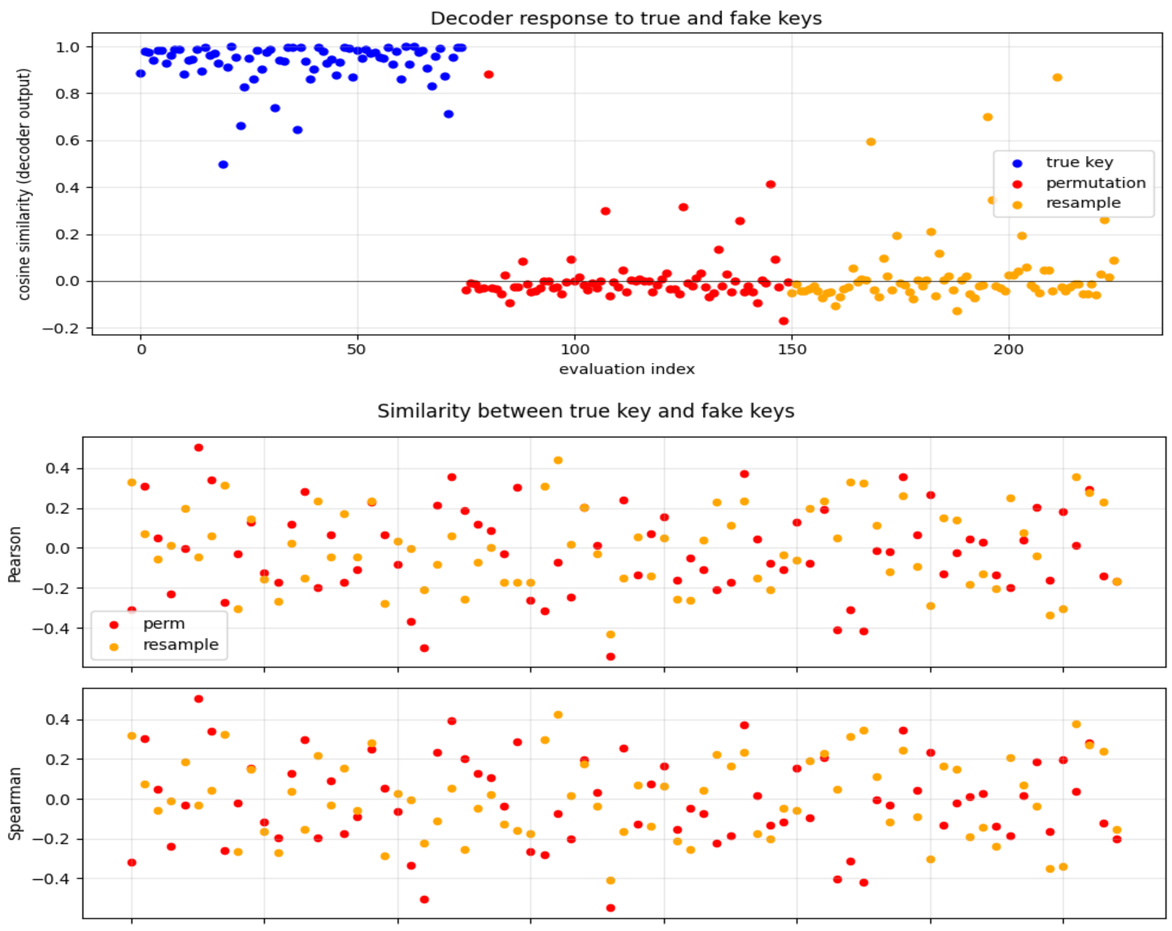
Transformation of hidden states via the decoder for geometric signature verification.

Training objective

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{tpl}\mathcal{L}_{tpl} + \lambda_{rank}\mathcal{L}_{rank} + \lambda_{kl}\mathcal{L}_{kl}$$

- **Perceptual losses:**
  - $\mathcal{L}_{ce}$: preserves the original LM objective.
  - $\mathcal{L}_{kl}$: aligns the watermarked distribution with the vanilla model.
- **Separation (template) loss** $\mathcal{L}_{tpl}$: preserves the geometric signal
- **Contrastive (ranking) loss** $\mathcal{L}_{rank}$: trains the decoder and enforces real key → high cosine similarity, fake/clean → low similarity.

Key-sensitivity evaluation



- Decoder outliers are random and not correlated. Forged keys produce no consistent similarity signal, indicating only true keys activate the decoder.

Key insights and future directions

Key insights

- Watermark embedded directly in the latent space via a geometric signature
- Exploits inherent RoPE properties used by many modern LLMs
- Triggered only by structured positional patterns

Future directions

- **Scaling:** evaluate on larger foundation models
- **Robustness evaluation** to fine-tuning, pruning and quantization attacks

elliot.cole@telecom-sudparis.eu mihai.mitrea@telecom-sudparis.eu hakima.chaouchi@imt.fr

This work is supported by BPI France SIMPLERAN Project https://www.simpleran.org/.