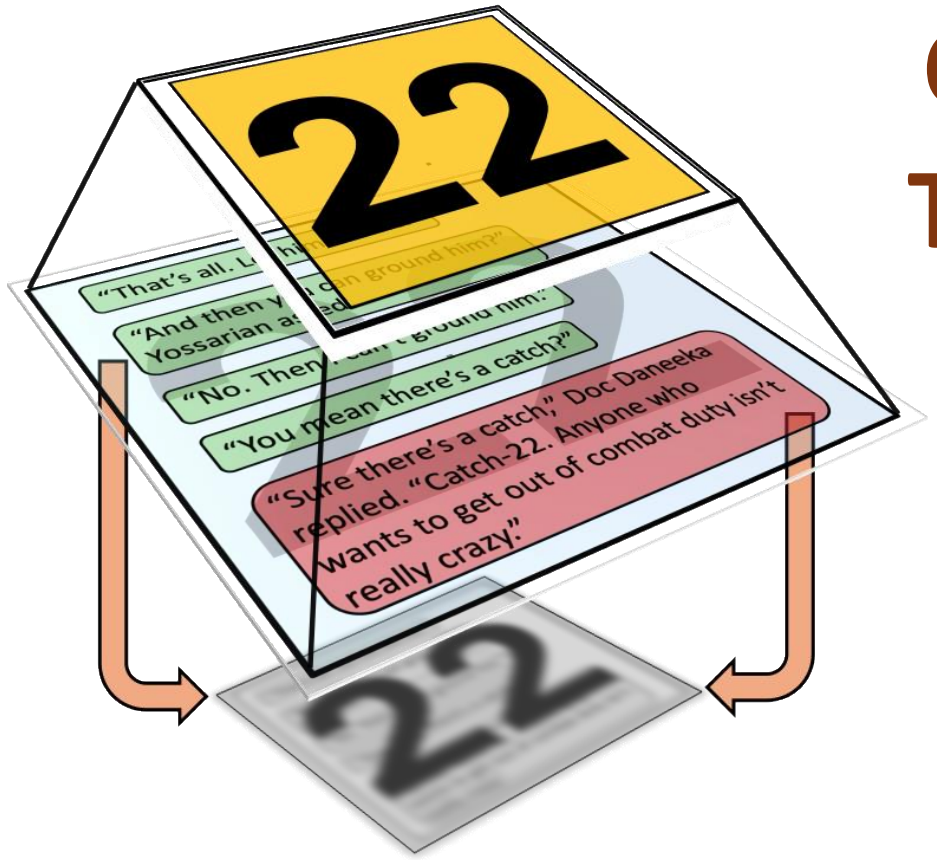
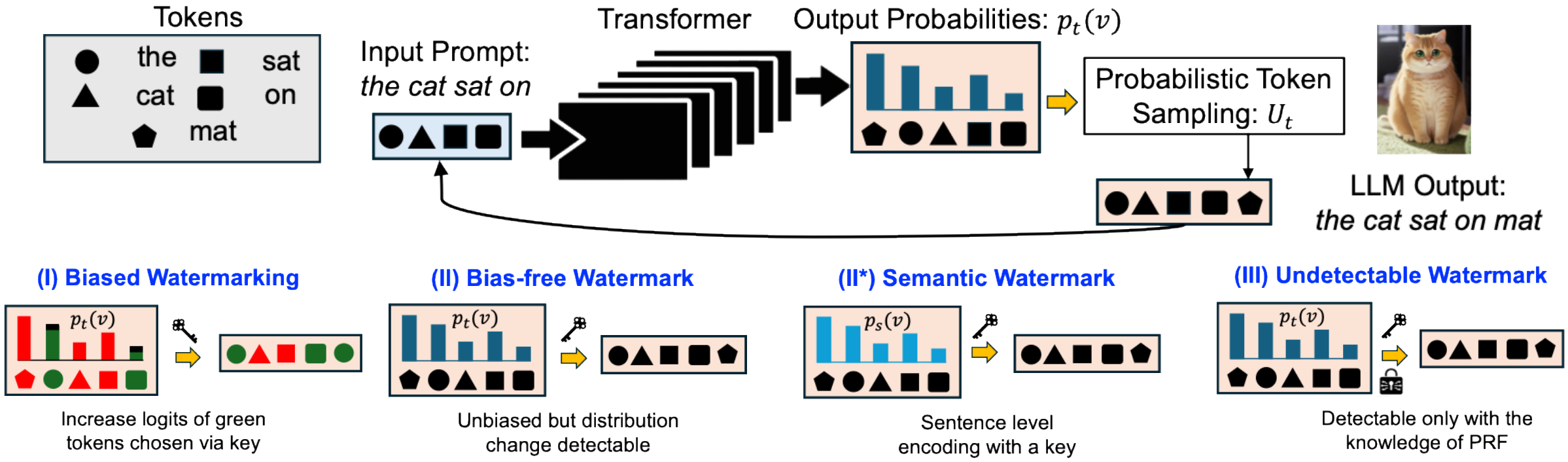


Catch-22: On the Fundamental Tradeoff Between Detectability and Robustness in LLM Watermarking

Kuheli Pratihar and Debdeep Mukhopadhyay,
Indian Institute of Technology Kharagpur, India

ICML
International Conference
On Machine Learning

Detectability-Robustness Tradeoff

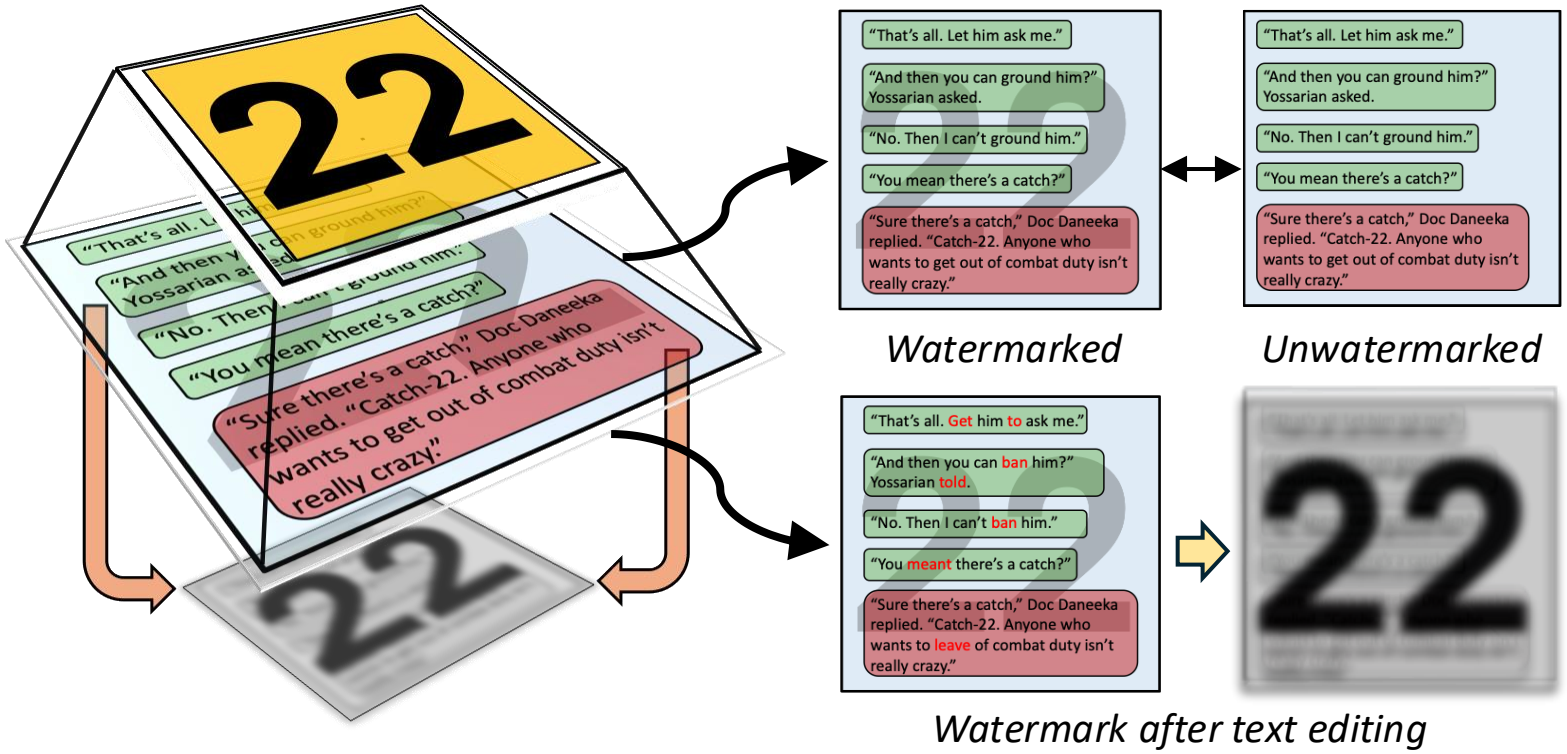


Since 2023 [Kirchenbauer et al.] watermark design research has developed significantly, but:

(i) lack a unified theoretical standpoint that can compare different watermark families

(ii) tie a scheme's robustness and keyless detectability directly to construction

Quantifying detectability & robustness across different watermarks



Sampling family (T output tokens)	Detectability (TV bound)
Greedy	$O(1)$
Biased (δ tilt)	$O(\lvert\delta\rvert\sqrt{\mathrm{T}})$
Bias-free	$O(\sqrt{\mathrm{T}})$
Semantic (avg. sentence length l)	$O(\sqrt{\mathrm{T}/l})$
Distribution-preserving	0

We quantify the upper bound on detectability in terms of total variation distance

Sampling family (T output tokens)	Post-edit KL budget	Maximum Tolerable edit-rate
Token-level (biased /bias-free)	$\mathrm{T}_s(1-\epsilon)^2\mathrm{D}_0^{(\mathrm{tok})}$	$1\text{-}\sqrt{\log_2(1/\beta)/\mathrm{TD}_0^{(\mathrm{tok})}}$
Semantic	$\mathrm{T}_s(1-2\epsilon_s)^2\mathrm{D}_0^{(\mathrm{sem})}$	$\leq \frac{1}{2}\left(1-\sqrt{\log_2(1/\beta)/\mathrm{TD}_0^{(\mathrm{sem})}}\right)$

Robustness of watermark after edits to the LLM output decay quadratically

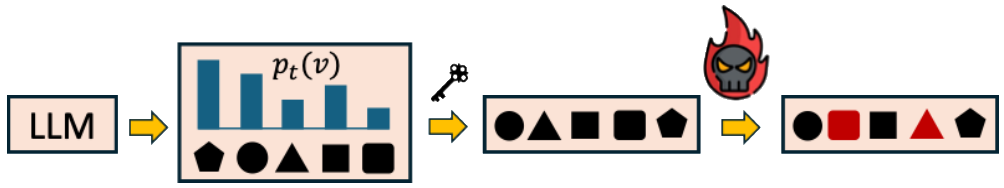
Theorem 1 [informal]: Any probability-modifying watermark leaves a statistical trace whose strength grows with the length of the generated text.

Theorem 2 [informal]: Watermark detection power contracts quadratically under text edits, imposing a fundamental trade-off between stealth and robustness.

Theoretical Contributions: Theorem 1 and 2 are formally defined in paper as Theorem 3.2 and 3.4.

Experiments across different watermarking schemes

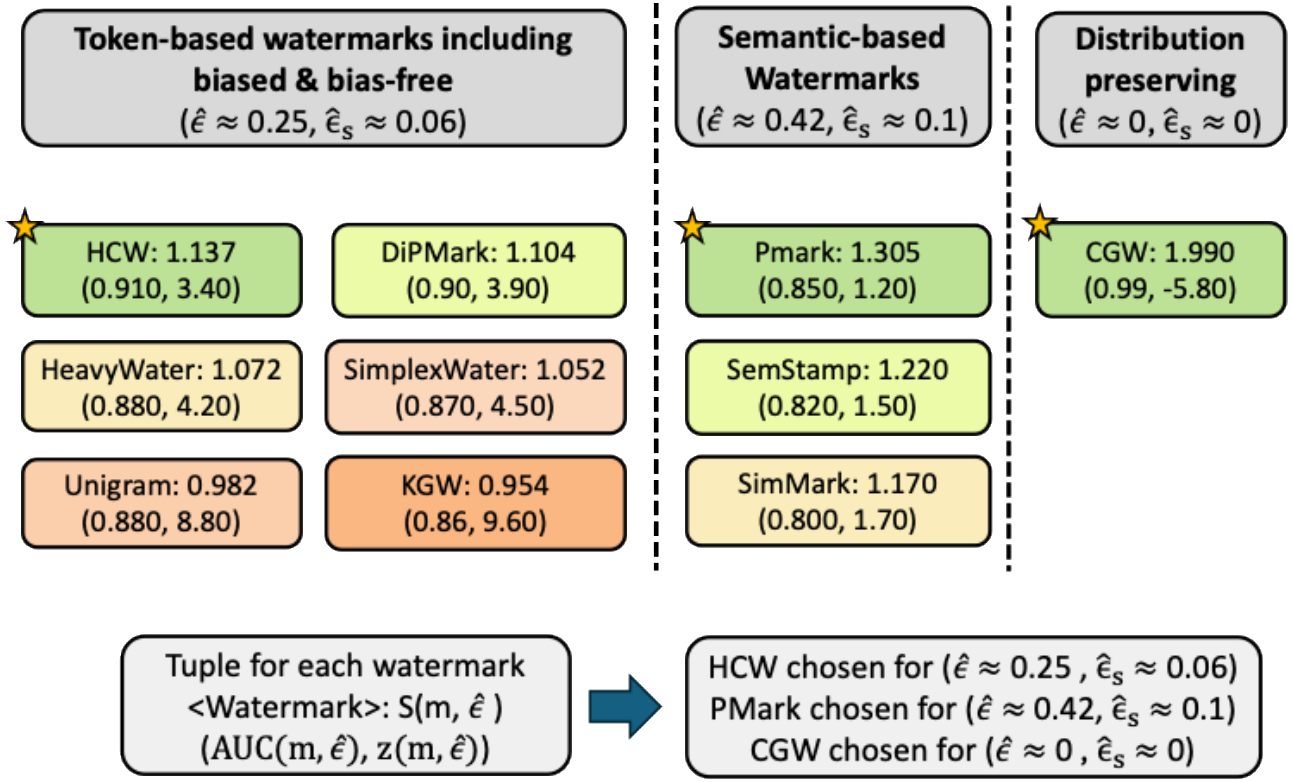
Theorem 3 [informal]: The hybrid selection rule provably picks the watermark scheme with the smallest stealth cost that still meets the target robustness under the anticipated edit regime.



Edit rate estimation -> ϵ ➜ hybrid selection rule

$$S(m) = \mathrm{AUC}(\mathrm{m}, \hat{\epsilon}) + 1/(1 + \max(\mathrm{z}(\mathrm{m}, \hat{\epsilon}), 0))$$

Score is computed for a target LLM at the estimated edit rate.



We empirically find that our proposed hybrid rule attains the highest score across 8 watermarking families on Llama-2 and Mistral-7B models

Thank you for visting!

You are cordially invited to my
Poster (**ID 66807**)

Wed, Jul 8, 2026

from 5:00 PM – 6:45 PM KST

**COEX Convention & Exhibition Center,
HALL A, Seoul, South Korea**

We are thankful for the partial funding from the Center for Hardware-Security Entrepreneurship Research & Development (C-HERD) and Information Security Education and Awareness (ISEA), both initiatives of MeitY, Govt. of India.