



Is Multilingual LLM Watermarking Truly Multilingual? Scaling Robustness to 100+ Languages via Back-Translation

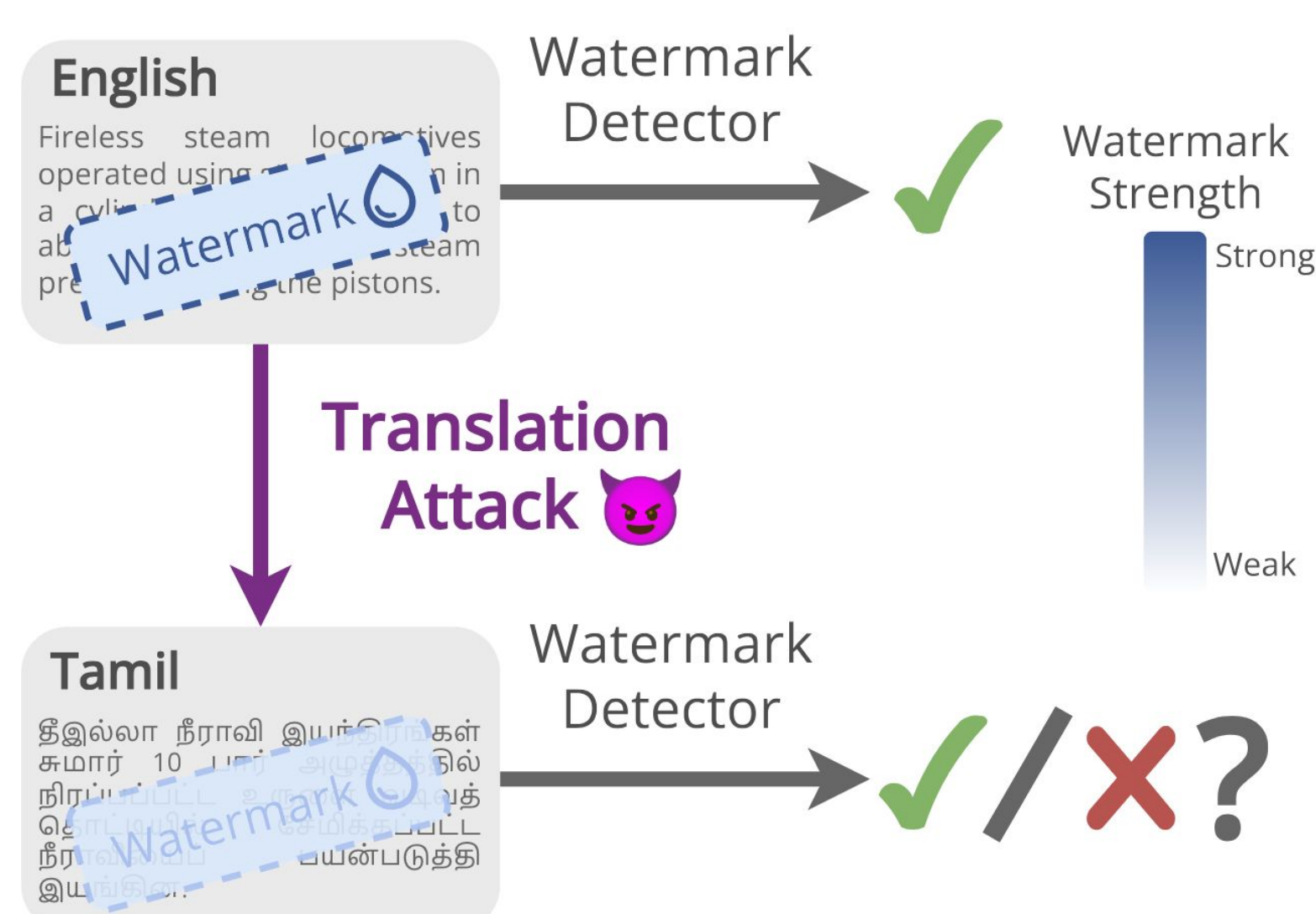
Asim Mohamed¹, Martin Gubri^{2 3}

¹African Institute for Mathematical Sciences, ²Parameter Lab, ³École Polytechnique, IP Paris, CNRS

We introduce STEAM for robust multilingual watermarking

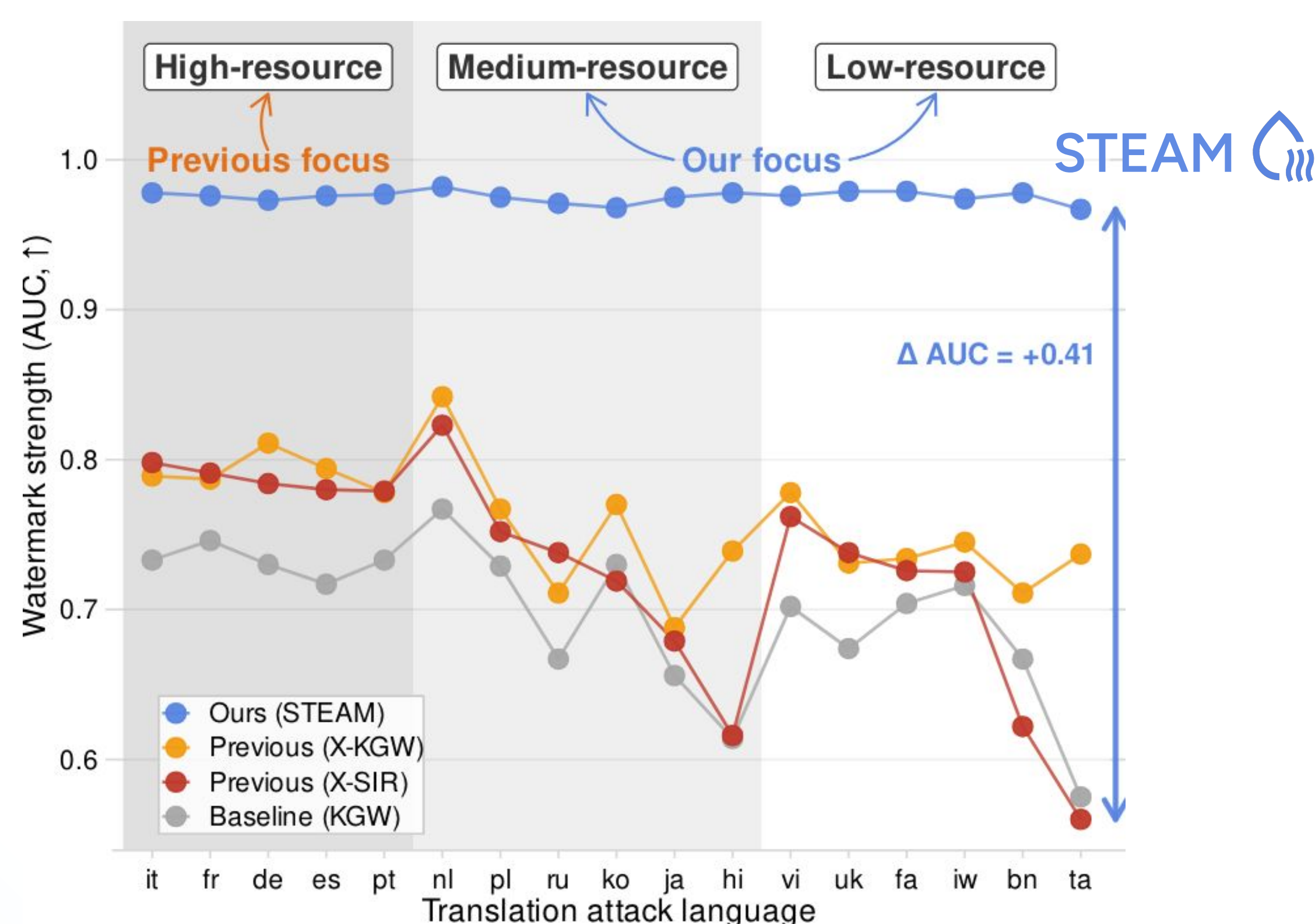
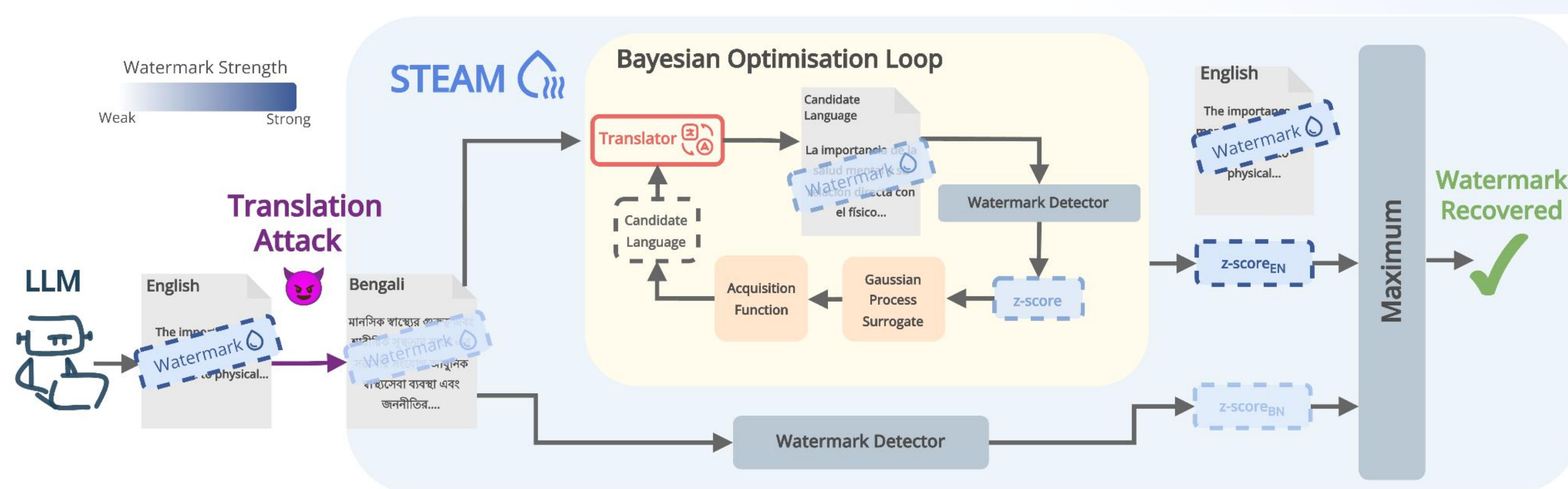
Overview

Multilingual watermarking: robustness to translation attack



STEAM

STEAM: Search-based Translation-Enhanced Approach for Multilingual watermarking



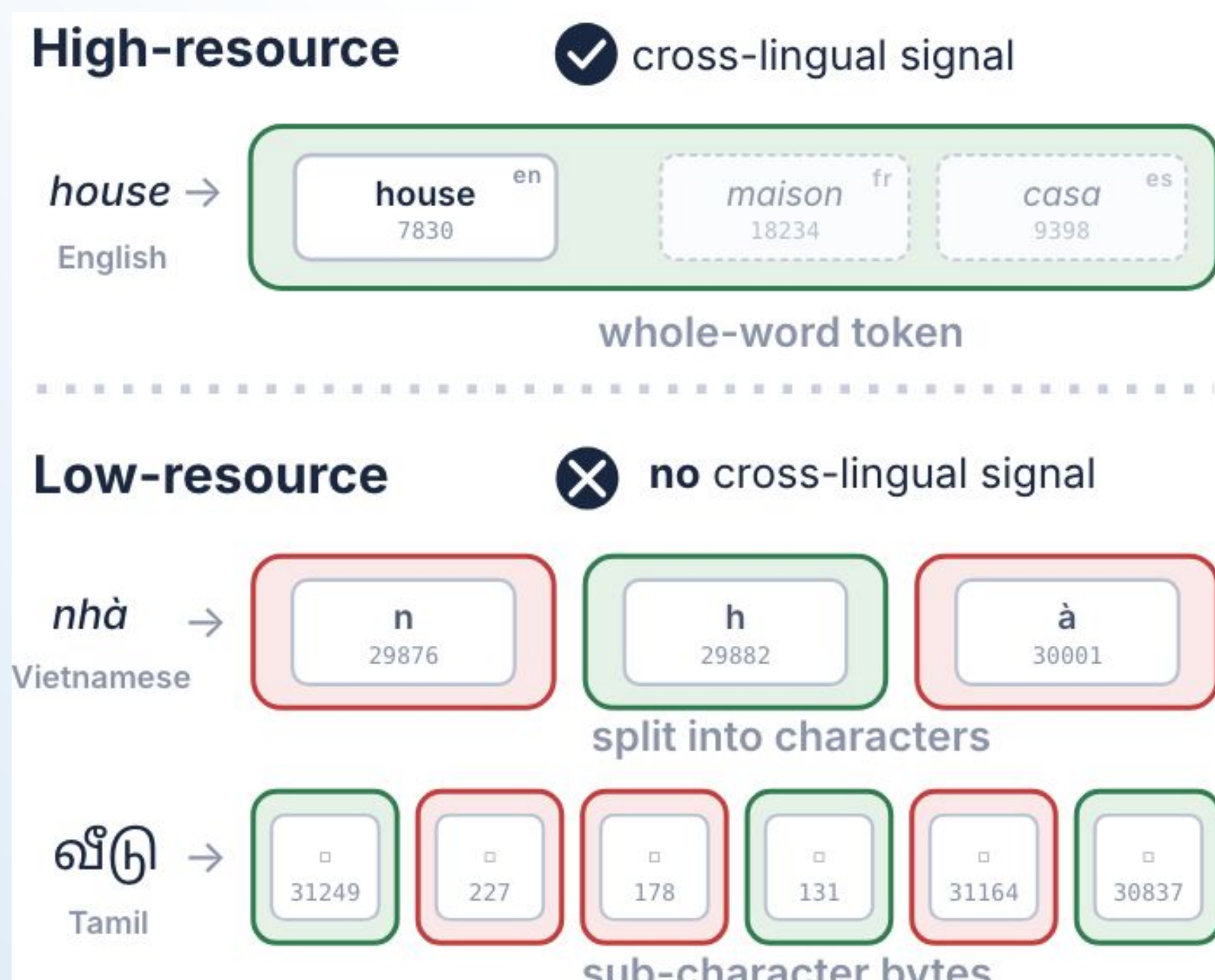
Comparison to previous work

Source Language		AUC (↑)				TPR@1% (↑)			
Type	Language	KGW	X-KGW	X-SIR	STEAM	KGW	X-KGW	X-SIR	STEAM
High-resource	fr	0.746	0.787	0.791	0.976	0.224	0.280	0.149	0.582
	de	0.730	0.811	0.784	0.973	0.224	0.312	0.163	0.622
	it	0.733	0.789	0.798	0.978	0.202	0.354	0.152	0.530
	es	0.717	0.794	0.780	0.976	0.232	0.278	0.150	0.580
	pt	0.733	0.778	0.779	0.977	0.246	0.330	0.176	0.538
Medium-resource	pl	0.729	0.767	0.752	0.975	0.228	0.312	0.146	0.526
	nl	0.767	0.842	0.823	0.982	0.290	0.332	0.213	0.612
	ru	0.667	0.711	0.738	0.971	0.158	0.246	0.122	0.576
	hi	0.614	0.739	0.616	0.978	0.120	0.194	0.056	0.644
	ko	0.730	0.770	0.719	0.968	0.210	0.318	0.115	0.460
Low-resource	ja	0.656	0.688	0.679	0.975	0.114	0.160	0.103	0.498
	bn	0.667	0.711	0.622	0.978	0.068	0.180	0.055	0.604
	fa	0.704	0.734	0.726	0.979	0.196	0.242	0.131	0.664
	vi	0.702	0.778	0.762	0.976	0.186	0.308	0.157	0.626
	iw	0.716	0.745	0.725	0.974	0.172	0.220	0.115	0.578
	uk	0.674	0.731	0.738	0.979	0.210	0.222	0.148	0.542
	ta	0.575	0.737	0.560	0.967	0.082	0.172	0.049	0.504

- Substantially better than semantic clustering
X-KGW: +0.22 AUC, +30.7%p TPR@1%
X-SIR: +0.25 AUC, +44.0%p TPR@1%
- Consistent across 126 languages
Average AUC of 0.976
Only 2 languages below 0.965
- Even with hold-out source language
Average AUC of 0.967

Previous Work

Semantic clustering (X-KGW, X-SIR) groups whole-word tokens. Not effective for languages with few/none whole-word tokens.



Hold-out languages

New Language	AUC (↑)			
	KGW	X-KGW	X-SIR	STEAM
it	0.733	0.772	0.796	0.966
es	0.717	0.807	0.754	0.967
pt	0.732	0.792	0.775	0.971
pl	0.730	0.762	0.749	0.960
nl	0.768	0.808	0.776	0.966
hr	0.706	0.757	0.726	0.965
cs	0.717	0.754	0.773	0.974
da	0.713	0.764	0.734	0.971
ko	0.732	0.754	0.729	0.961
ar	0.689	0.765	0.687	0.971

False Positive Analysis

Budget	Pool Size	FPR ($z=2.33$) (↓)	τ^*
Fixed	33	29.2%	3.94
	66	25.8%	3.87
	126	25.8%	3.90
Proportional	33	17.7%	3.60
	66	25.8%	3.87
	126	41.2%	4.07

More languages with constant search budget does not inflate false positive rate

Robustness Analysis

Translators mismatch

Attacker	Defender		
	G. Translate	DS-V3.2	GPT-4o-mini
Google Translate	0.976	0.974	0.975
DS-V3.2	0.979	0.952	0.943
GPT-4o-mini	0.976	0.973	0.964

Also robust for:

- Other source languages
- Adaptive oracle attacker (worst language)
- Paraphrase-then-translate attack
- Multi-step translation attack
- Local open-weights translators (NLLB)

Ablation Study

Language-specific null hypothesis with normalized z-score per language:

$$z_\ell = (n_g - \gamma_\ell n) / \sqrt{n \gamma_\ell (1 - \gamma_\ell)}$$

candidate language number of green tokens fraction of green tokens on 500 hold-out human texts

Z-score normalization	Watermark		Language Pred.
	AUC ↑	TPR@1% ↑	Accuracy (%) ↑
No	0.902 ±0.11	0.34 ±0.13	38.6 ±16.6
Yes (STEAM)	0.951 ±0.03	0.64 ±0.17	83.5 ±09.0

Conclusion

Previous multilingual watermarks are not multilingual. STEAM improves detection beyond high-resource languages.

