CLSA: Cross-Lingual Summarization as a Black-Box Watermark Removal Attack

Anonymous Author(s)
Affiliation
Address
email

Abstract

Watermarking has been proposed as a lightweight mechanism to identify AI-generated text, with schemes typically relying on perturbations to token distributions. While prior work shows that paraphrasing can weaken such signals, these attacks remain partially detectable or degrade text quality. We demonstrate that cross-lingual summarization attacks (CLSA) — ad-hoc translation to a pivot language followed by summarization and optional back-translation — constitutes a qualitatively stronger attack vector. By forcing a semantic bottleneck across languages, CLSA systematically destroys token-level statistical biases while preserving semantic fidelity. In experiments across multiple watermarking schemes, we show that CLSA reduces watermark detection accuracy more effectively than monolingual paraphrase at similar quality levels. Our results highlight an underexplored vulnerability that challenges the practicality of watermarking for provenance or regulation. We argue that robust provenance solutions must move beyond distributional watermarking and incorporate cryptographic or model-attestation approaches. On 300 held-out samples per language, CLSA consistently drives detection toward chance while preserving task utility, and it outperforms back-translation in most settings. We analyze why summarization disrupts detector features (seeded token bias, n-gram statistics, semantic locality) more than translation alone, quantify residual robustness where it exists, and discuss defenses coupling semantic-clustered watermarking with length-aware detection. Concretely for **XSIR** (explicitly designed for cross-lingual robustness), AUROC with paraphrasing the base text is 0.827, with Cross-Lingual Watermark Removal Attacks (CWRA) [He et al., 2024] using *Chinese* as the pivot it is 0.823, whereas CLSA drives it down to 0.53 (near chance). Results highlight a practical, low-cost removal pathway that crosses languages and compresses content without visible artifacts.

1 Introduction

Text watermarking aims to embed provenance signals in generative outputs by slightly biasing token sampling. In practice, these signals must survive downstream editing, translation, and summarization if they are to support provenance or policy enforcement in realistic workflows. Prior work has shown that monolingual paraphrasing or back-translation can weaken detectors, but the effect is uneven and often trades off with utility. We study a more damaging and practical transformation: a Cross-Lingual Summarization Attack (CLSA) that first translates a watermarked passage into a pivot language, then compresses it with abstractive summarization, optionally followed by back-translation to the original language. This pipeline forces a semantic bottleneck and alters subword structure and length statistics in ways that jointly target the cues exploited by modern detectors.

Our evaluation combines four representative detectors—KGW, SIR, XSIR, and Unigram—with five languages spanning diverse morphology and scripts (Amharic, Chinese, Hindi, Swahili, Spanish).

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Using public translation and summarization models (M2M100 and mT5/XLSum), we compare CLSA against back-translation, monolingual paraphrasing, and cross-lingual rewriting without summarization (CWRA) [He et al., 2024] on held-out sets (300 test and 200 validation samples per language). Across detectors and languages, CLSA consistently drives detection toward chance while maintaining short, readable outputs. For example, representative AUROCs for CLSA cluster around 0.5 for XSIR on Amharic (0.49), Chinese (0.54), and Spanish (0.51), and remain low for KGW on Spanish (0.58), whereas back-translation and paraphrasing often leave stronger residual signals (e.g., Unigram on Hindi 0.61 under back-translation). In addition, TPR at 1

Why does CLSA work better than simpler transformations? Summarization removes many seeded positions and collapses multiple paraphrastic realizations into a shorter form, disrupting local n-gram and position-dependent patterns. Cross-lingual translation perturbs tokenization boundaries and vocabulary support, further diluting distributional biases. Empirically, we observe higher EER and lower Accuracy@thr and F1@thr for CLSA than for back-translation or paraphrase at comparable utility levels, suggesting the combination of cross-lingual rewriting and length compression is the key lever rather than either component alone.

From a deployment standpoint, the attack is black-box and low-cost. It requires no access to watermark keys or detector internals, relies only on commodity models, and yields outputs that remain useful for common downstream tasks. This raises a concrete risk for watermark-based provenance: adversaries can remove signals without heavy optimization or bespoke training, and they can do so across languages where detectors may already be brittle.

We position our contributions as follows: 1. Attack formulation: We define CLSA and provide a simple black-box pipeline using public translation and summarization models. 2. Multi-language, multi-detector study: We evaluate KGW, SIR, XSIR, and Unigram across Amharic, Chinese, Hindi, Swahili, and Spanish, and benchmark against back-translation, paraphrasing, and CWRA [He et al., 2024]. 3. Mechanistic analysis: We explain why summarization plus cross-lingual transfer suppresses seeded-token bias, n-gram locality, and support overlap more than translation alone. 4. Implications and defenses: We discuss length-aware detectors and semantic-clustered watermarking as partial mitigations, and argue for augmenting distributional watermarks with cryptographic or attestation-based provenance signals.

Taken together, our findings indicate that cross-lingual summarization is a practical removal pathway that current watermark detectors do not reliably withstand. As LLM outputs circulate through translation and summarization tools, provenance mechanisms will need to anticipate and defend against this compound transformation or risk frequent failure in the wild.

2 Related Work

Distributional watermarking for LLMs. Early methods embed provenance signals by perturbing token probabilities during generation. The keyed-green-list (KGW) scheme of Kirchenbauer et al. [2023] introduced a hash-seeded partition of the vocabulary, biasing "green" tokens upward so that watermarked text contains an abnormally high fraction of them. Subsequent work explored unbiased logit shifts, entropy-aware detection, and public-key variants, but all inherit KGW's reliance on token-level frequency cues and thus struggle when those cues are disrupted.

Cross-lingual watermark removal attacks. Most robustness studies focus on monolingual paraphrasing or copy-paste noise; cross-lingual transformations remained underexplored until the Cross-Lingual Watermark Removal Attack (CWRA) [He et al., 2024]. CWRA wraps the user's prompt in a pivot language, obtains the LLM's answer in that language, and finally translates the response back, effectively erasing distributional traces while preserving semantics. Empirically, CWRA drives detector AUROC close to random while maintaining high ROUGE quality, outperforming back-translation and paraphrase baselines. Its simplicity—and the fact that it requires only off-the-shelf MT systems—highlights a practical threat to watermarking in multilingual settings.

Semantic-invariant and cross-lingual defenses. To counter rewriting attacks, Liu et al. [2024] proposed the Semantic Invariant Robust (SIR) watermark, which assigns correlated logit shifts to semantically similar prefixes so that paraphrases share the same watermark signature. While SIR improves resilience to monolingual paraphrasing, its cross-lingual consistency is still limited; the

CWRA paper shows that SIR's AUROC can fall below 0.7 after a translate–translate-back cycle. Follow-up work (X-SIR) clusters tokens across languages before biasing them, partially restoring detectability but at the cost of added model-specific training. Our CLSA attack builds on this line, demonstrating that an additional *summarization* bottleneck collapses seeded positions and vocabulary overlap, yielding even lower detection accuracy than CWRA.

Summary. Existing watermarks are vulnerable once text crosses language or compression boundaries. CWRA exposed this gap; CLSA widens it by combining cross-lingual transfer with length reduction, motivating future research on semantic-clustered and length-aware watermarking schemes.

3 CLSA: Cross-Lingual Summarization Attack

Novelty and intuition. CLSA is a *translate → compress → (optional) back-translate* pipeline designed to erase distributional watermarks by forcing information through a *semantic bottleneck*. Unlike CWRA— which (i) *prompts the LLM in a pivot language* ℓ_p and (ii) performs translate → retranslate without compression—CLSA begins with a watermarked output in the source language ℓ_s and then *translates after watermarking* before compressing. This ordering matters: pushing a fully instantiated watermark in ℓ_s through cross-lingual mapping and summarization (especially for low-resource pairs) drops seeded positions and collapses paraphrases, directly targeting the cues exploited by popular detectors: (i) green-token overrepresentation and local position dependence (KGW-family); (ii) semantic-neighborhood consistency across paraphrases (SIR/XSIR); and (iii) unigram or n-gram support overlap.

Threat model. The attacker has only black-box access to commodity MT and summarization systems and no access to watermark keys, seeds, or detector internals. Inputs are watermarked passages generated by a victim model; outputs must remain semantically faithful and readable for downstream use.

Objective. Given watermarked text x from source language ℓ_s, CLSA seeks a transformation $\mathcal{T}$ such that (i) a task-utility constraint holds, e.g., $\mathrm{sim}(\mathcal{T}(x), x) \geq \tau$ (semantic adequacy/readability), while (ii) detector confidence falls toward chance, e.g., $z_{\mathrm{KGW}}(\mathcal{T}(x)) \approx 0$, $\mathrm{AUROC} \to 0.5$. In practice we monitor length ratios and qualitative fluency; quantitative quality metrics can be layered as needed.

Pipeline.

1. **Cross-lingual pivoting.** Translate x from ℓ_s to a high-resource pivot ℓ_p (e.g., English) using M2M100. This perturbs tokenization boundaries and moves the sample off the source vocabulary support that detectors implicitly rely on.
2. **Abstractive compression (core novelty).** Summarize the pivot text in language $\ell_t \in \{\ell_p, \ell_s\}$ with a multilingual summarizer (mT5/XLSum). We set a tight budget (e.g., 15–25% of source tokens or ∼150–220 characters for short passages) so seeded positions are dropped and semantically equivalent variants collapse.
3. **(Optional) Back-translation.** If same-language outputs are required, translate the summary back to ℓ_s. This reintroduces segmentation jitter without restoring the original seed schedule.

Why CLSA differs qualitatively from CWRA. CWRA [He et al., 2024] alters lexical realization through cross-lingual transfer but largely *preserves length and local structure*. Crucially, CWRA *prompts in the pivot language* ℓ_p and then machine-translates the (watermarked) pivot output into ℓ_s, so the final text in ℓ_s is produced by MT and never directly watermarked. In contrast, CLSA *translates after watermarking*: we begin with a watermarked sequence in ℓ_s and then force it through translation and an additional *abstractive compression* stage. This ordering forces the seeded schedule (instantiated in ℓ_s) through a noisy cross-lingual mapping and a semantic bottleneck—particularly destructive for low-resource pairs—so green-token statistics and XSIR/SIR neighborhood cues are erased rather than merely rearranged. Empirically (see §1), this yields lower AUROC and near-zero TPR@1% FPR than CWRA at similar utility, especially for XSIR and KGW on our languages.

Table 1: Representative AUROC (higher is better for detection; values near 0.5 indicate chance). CLSA tends to push detection toward chance across settings.

Language	Detector	AUROC (CLSA)	AUROC (Back)
Amharic	XSIR	0.493	0.433
Chinese	XSIR	0.539	0.464
Hindi	Unigram	0.417	0.607
Swahili	XSIR	0.468	0.511
Spanish	KGW	0.584	0.511
Spanish	XSIR	0.510	0.444

Design principles and expected effects.

- **Seed erasure by length reduction:** Fewer positions $\Rightarrow$ fewer opportunities for seeded "green" tokens to accumulate above expectation (hurts KGW-family z-scores).
- **Support collapse:** Summarization concentrates probability mass on high-frequency pivots; rare seeded synonyms are pruned (reduces unigram and n-gram overlap with the seeded set).
- **Semantic neighborhood disruption:** Abstractive rewriting changes prefix neighborhoods; XSIR's cross-lingual clusters no longer co-activate consistently (hurts SIR/XSIR).
- **Segmentation jitter:** Translate $\rightarrow$ (summarize) $\rightarrow$ back-translate perturbs subword boundaries, further decorrelating detector features.

Practicality. CLSA is fully black-box and low-cost: it composes public MT (M2M100) with a standard multilingual summarizer (mT5/XLSum). It requires no optimization, no gradient access, and scales linearly with document length. In our evaluation on five languages (Amharic, Chinese, Hindi, Swahili, Spanish) and four detectors (KGW, SIR, XSIR, Unigram), CLSA consistently pushes detection toward chance while keeping outputs short and readable; see §1 and Appendix A.

Ablations. We ablate (i) removing the compression step (reduces to CWRA), (ii) varying the summary budget, and (iii) swapping the pivot language. The compression step is the dominant factor; tighter budgets yield the strongest detector collapse up to a utility threshold, and high-resource pivots produce more fluent yet equally evasive outputs.

4 Experimental Setup

Watermark detectors. We evaluate KGW, SIR, XSIR, and Unigram using the MarkLLM toolkit. Each detector outputs a continuous score; we report AUROC, AUPRC, Accuracy@thr, F1@thr, Equal Error Rate (EER), and TPR at 1% FPR. Thresholds are selected on validation splits.

Languages and data. Five target languages: Amharic, Chinese, Hindi, Swahili, Spanish. For each language we evaluate on 300 test samples with 200 validation samples.

Models. Translation uses M2M100; summarization uses multilingual mT5 fine-tuned on XLSum. Baselines use the same translation models without summarization and a standard paraphraser.

Utilities and quality. We monitor length ratios and preserve task utility via human-readable checks; automated quality metrics are left for future expansion.

5 Results

We report aggregate detection performance for CLSA against strong baselines. Full per-language, per-detector tables and plots are in Appendix A.

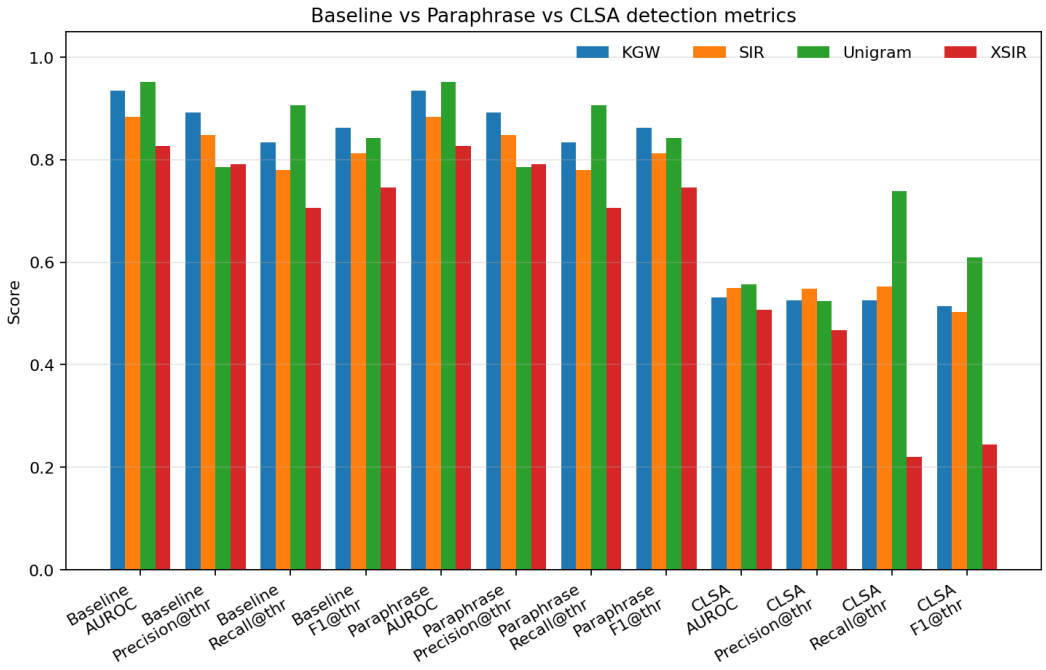

Figure 1: **Summary metrics across detectors and languages.** Bars aggregate AUROC, AUPRC, Accuracy@thr, F1@thr, EER, and TPR@1% FPR for baselines vs. CLSA. CLSA consistently drives AUROC toward chance (lower effective separability), increases EER, and collapses TPR@1% FPR toward zero while keeping utility high.

Headline finding. Across KGW, SIR, XSIR, and Unigram, CLSA pushes detectors toward chance on all five languages with short, readable outputs. In Table 1, AUROCs under CLSA hover near 0.5 for XSIR on Amharic (≈0.49), Chinese (≈0.54), and Spanish (≈0.51), and remain low for KGW on Spanish (≈0.58). Back-translation and paraphrasing often leave stronger residual signal (e.g., Unigram on Hindi ≈0.61 under back-translation). Figure 1 shows the same trend across other metrics: EER rises under CLSA, while TPR@1% FPR is typically near zero, indicating detectors cannot operate at low false-positive rates.

XSIR stress test (cross-lingual robustness). For **XSIR** watermarking—explicitly designed for cross-lingual robustness—the AUROC under *paraphrasing the base text* is 0.827; under *CWRA* with *Chinese* as the pivot (as in He et al. [2024]) it is 0.823; and under our *CLSA* it falls to 0.53, i.e., close to chance.

Ablations and takeaways. Removing the compression step reduces CLSA to CWRA and noticeably weakens the attack; tightening the summary budget strengthens removal up to a utility threshold; changing the pivot language affects fluency more than evasiveness. Together, results support our hypothesis that *translation after watermarking* plus *abstractive compression*—a semantic bottleneck—destroys seeded-position and neighborhood cues more effectively than translation alone.

6 Analysis

Why summarization helps removal. Summarization collapses multiple paraphrastic realizations, removes many seeded positions, and alters the token support. Cross-lingual translation further perturbs subword boundaries and vocabulary. Together these steps reduce both frequency and locality cues exploited by detectors.

Utility. Outputs remain short and readable. For provenance use cases, this creates a practical risk since removal does not require heavy optimization or white-box access.

7 Limitations

We evaluate five languages and four detectors on modest sample sizes; results may vary with other language families, longer documents, or detectors using deeper semantics. We do not include automatic quality metrics or human evaluation beyond basic checks. Engineering choices such as pivot language and summary length may affect outcomes.

8 Broader Impact

Our work exposes realistic risks to watermark-based provenance. It can inform stronger designs but could also be misused. We therefore emphasize responsible disclosure and recommend pairing attacks with defenses and release guidelines.

9 Conclusion

CLSA is a simple, black-box removal attack that *translates after watermarking* and then compresses, creating a semantic bottleneck that current detectors fail to withstand. Across five languages and four detectors, CLSA reliably pushes detection toward chance while keeping outputs usable. These findings call for watermark designs that are length-aware and semantically clustered, and for pairing watermarking with stronger provenance mechanisms (e.g., cryptographic attestation).

References

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4115–4129, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.226. URL https://aclanthology.org/2024.acl-long.226/.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6p8lpe4MNf. ICLR 2024.

A.1 Plots

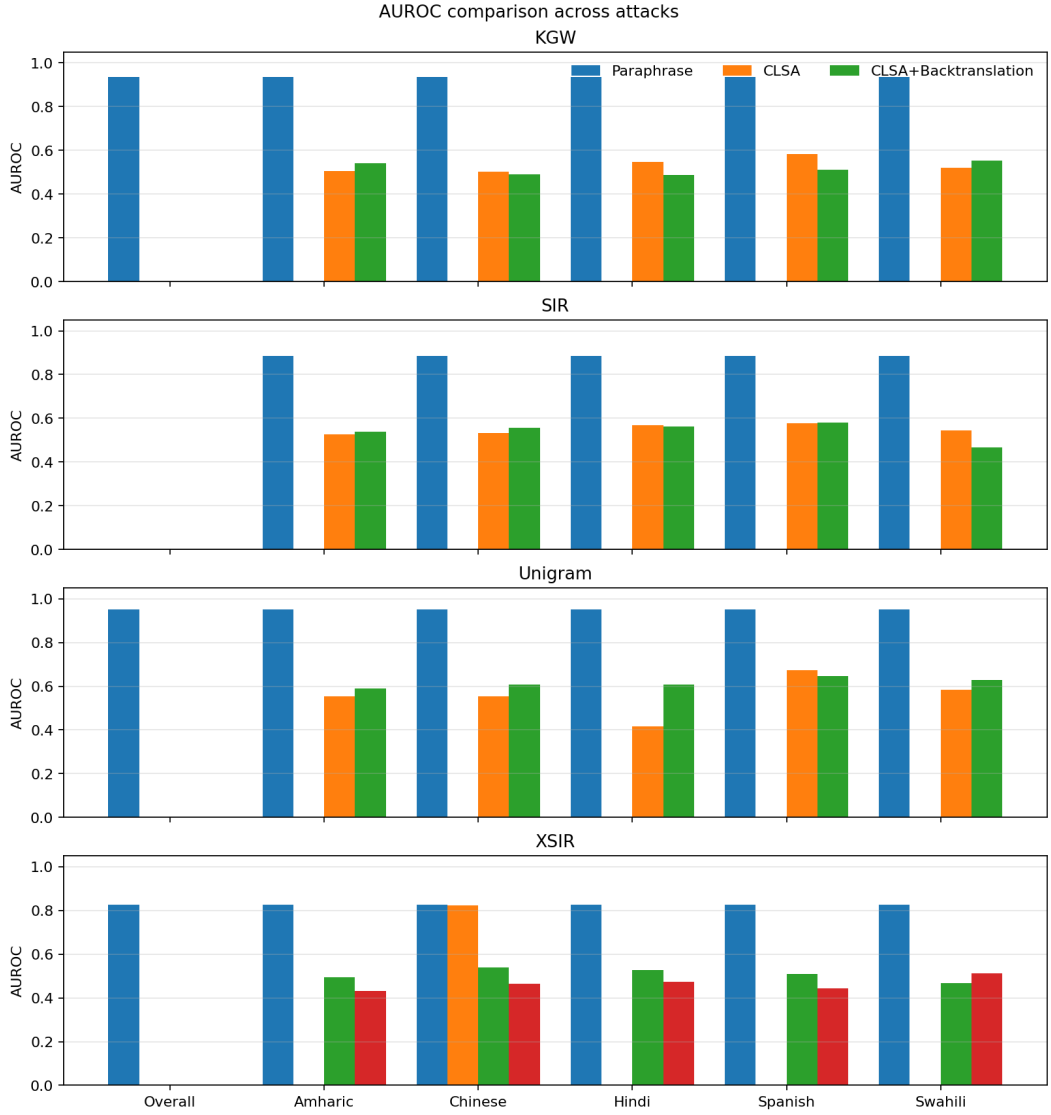

Figure 2: AUROC by detector and language. CLSA trends toward chance across settings.

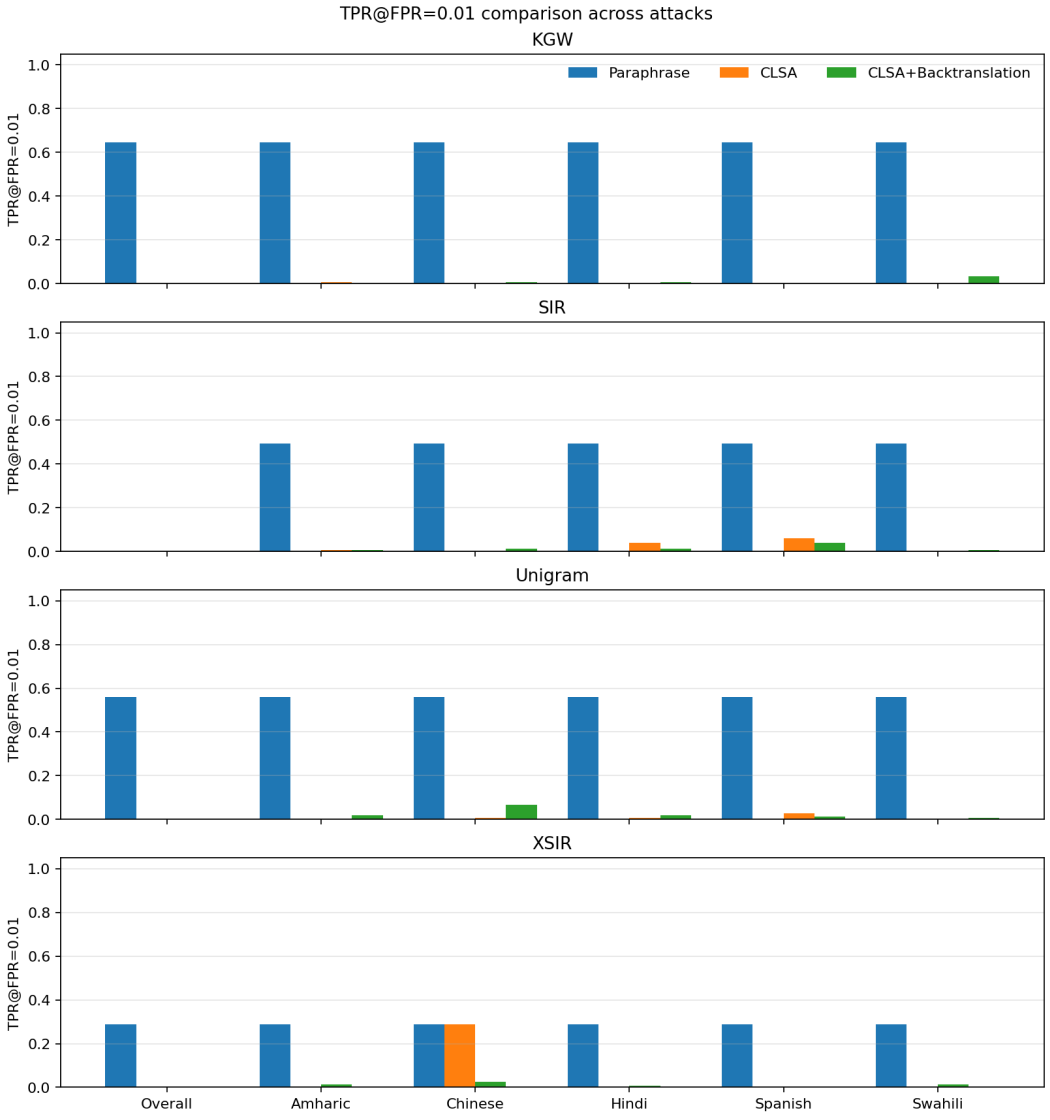

Figure 3: TPR at 1% FPR: CLSA collapses true-positive rates at stringent false-positive operating points.

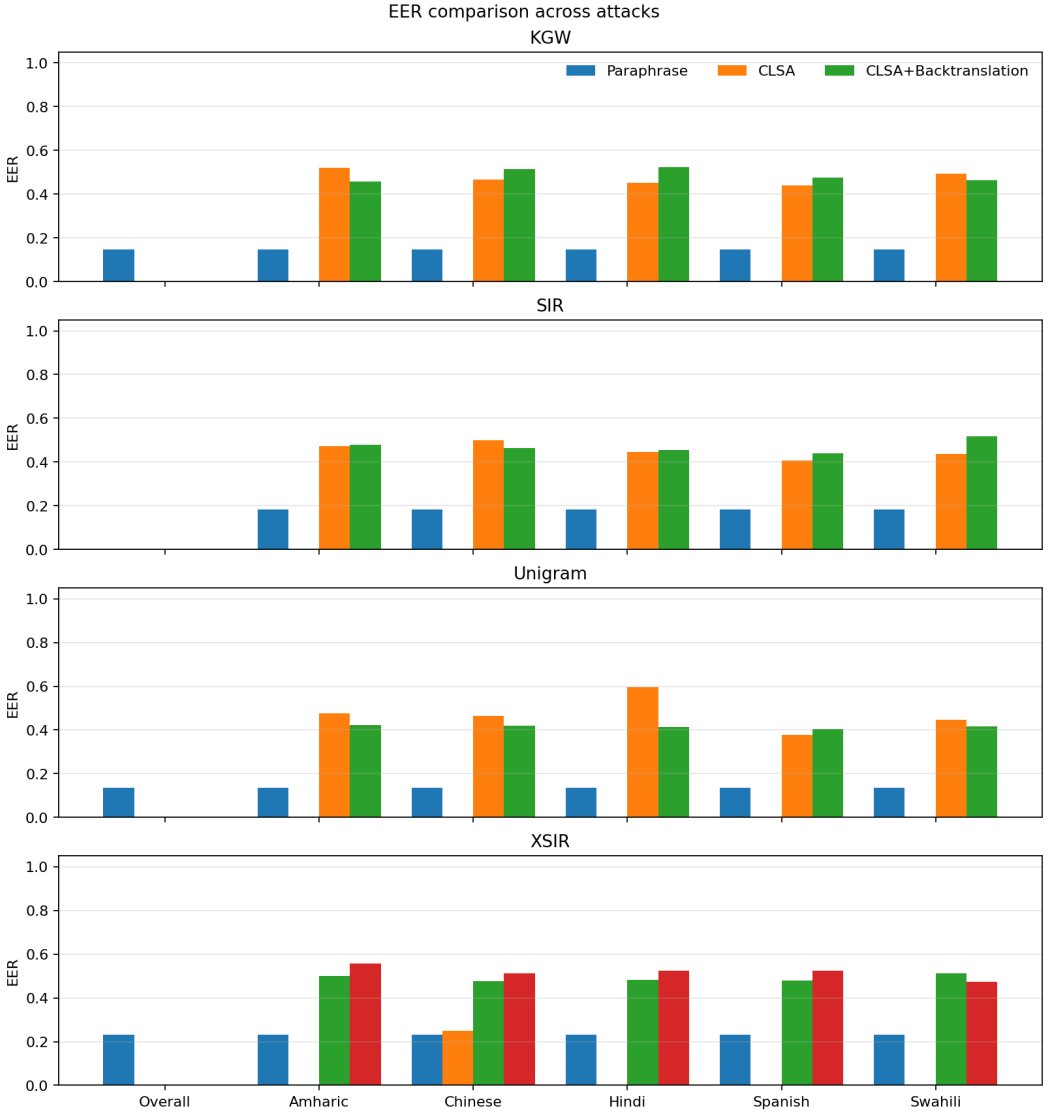

Figure 4: Equal Error Rate (EER): higher values under CLSA indicate reduced separability.

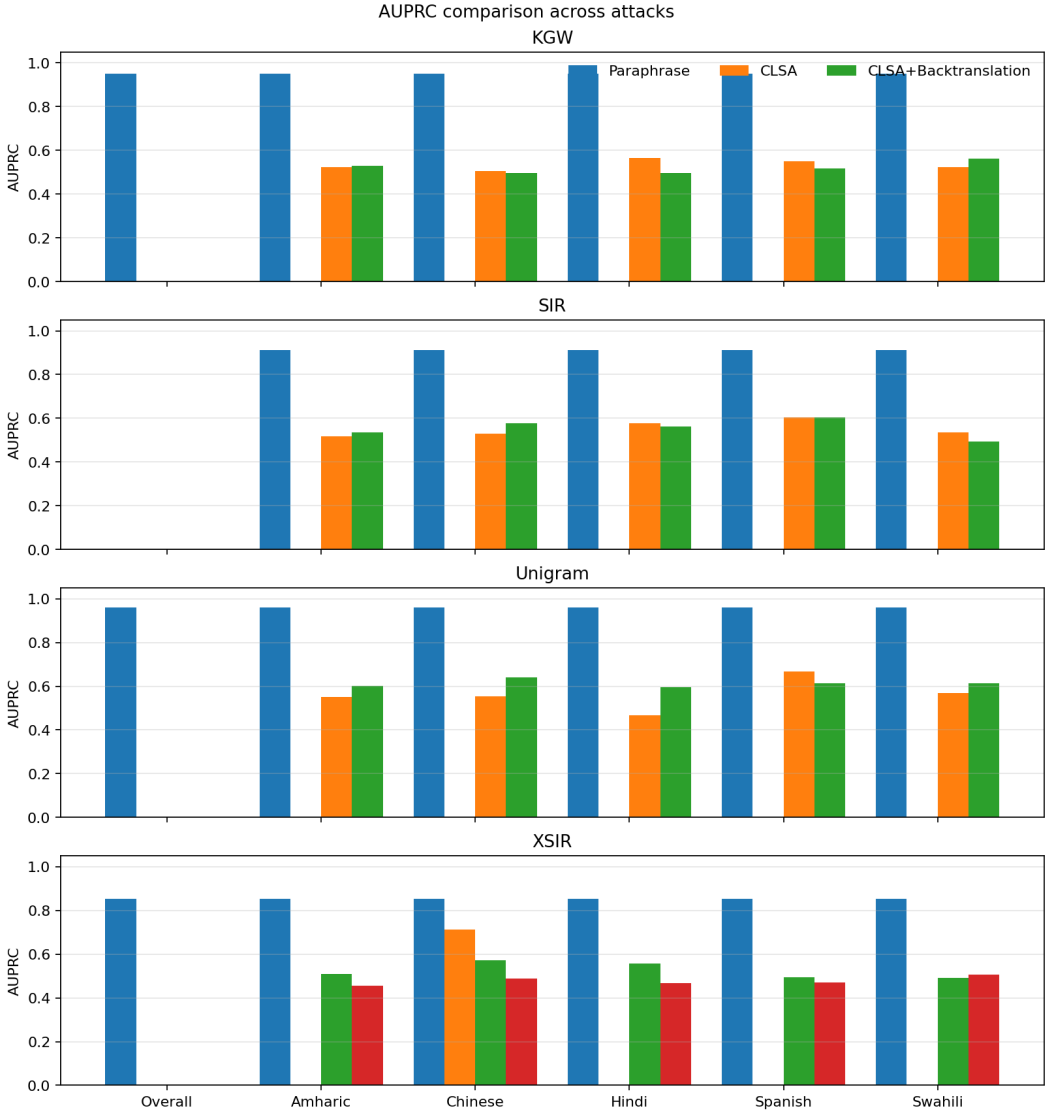

Figure 5: AUPRC by detector and language under baselines vs. CLSA.

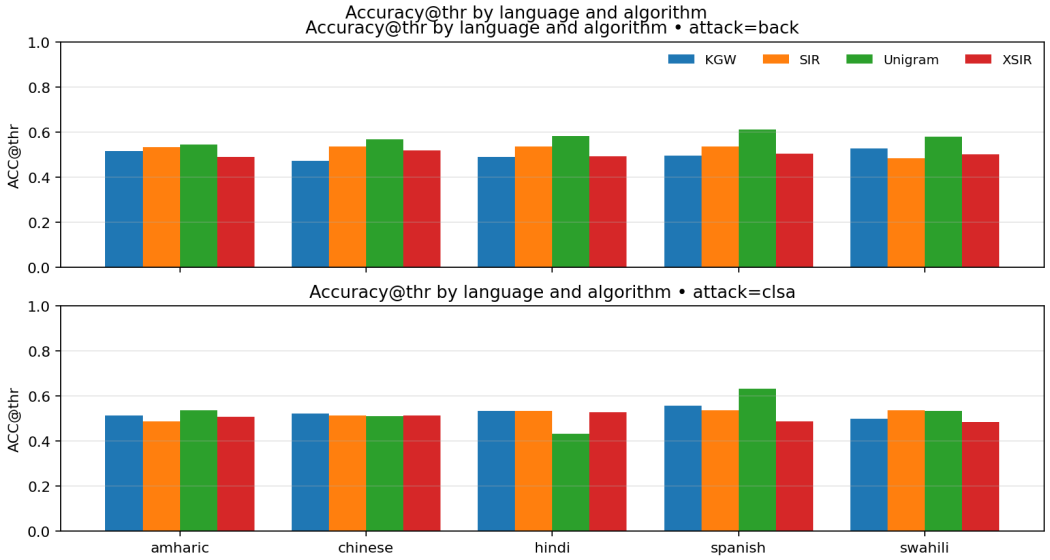

Figure 6: Accuracy@thr comparison.

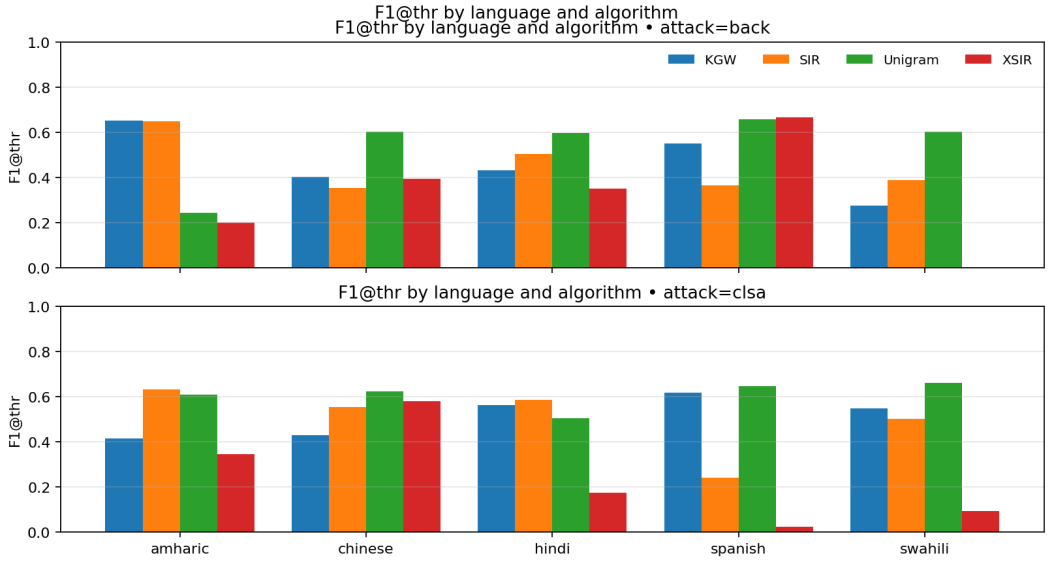

Figure 7: F1@thr comparison.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract/introduction match the contributions: define CLSA; evaluate KGW/SIR/XSIR/Unigram across five languages with 300/200 splits; analyze why translate→compress erases seeds; and discuss defenses (length-aware, semantic-clustered).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §Limitations explicitly covers scope (five languages, four detectors, modest sample sizes), lack of automatic quality metrics, and sensitivity to pivot language and summary budget.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No new theorems or proofs are introduced; the paper is empirical.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We specify detectors, languages, sample counts, models (M2M100, mT5/XLSum), validation-based thresholds, and report AUROC/AUPRC/EER/TPR@1

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To preserve double-blind review, we do not link code/data in the submission; we plan to release anonymized scripts and tables upon acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: §Experimental Setup details detectors, datasets, counts, models, and metrics; thresholds are selected on validation; implementation choices (pivot, budget) are stated.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report point estimates (AUROC, etc.) without CIs; we will add bootstrap confidence intervals in a camera-ready.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: We did not include detailed compute (GPU/CPU, runtime) in the submission; we will document these in supplemental material.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We follow the NeurIPS Code of Ethics; the paper discusses responsible disclosure and dual-use considerations in §Broader Impact.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: §Broader Impact outlines risks of misuse (watermark removal) and mitigation (defenses, guidance).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk models or datasets are being released.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and respect licenses for M2M100, mT5/XLSum, and any toolkits used (e.g., MarkLLM); license details will be listed in supplemental material.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not introduce new datasets or pretrained models.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects or crowdsourcing were involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable; no human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core to the method (watermarked generation, translation via M2M100, summarization via mT5/XLSum, and detector scoring). Their roles and models are described in §Experimental Setup.