Machine-Translated Bengali in the Wild: A Provenance Benchmark and a Detector-Flagged Shift Toward Translation in Common Crawl

Mantaqa Abedin
BRAC University
Dhaka, Bangladesh
mantaqa.abedin@g.bracu.ac.bd

Jeba Afia Khan
BRAC University
Dhaka, Bangladesh
jeba.afia.khan@g.bracu.ac.bd

Abstract

Low-resource Asian machine translation draws on monolingual web text from Common Crawl. We ask what share of Bengali Common Crawl a detector flags as machine output, and whether it moves with translation or generation. We introduce BanglaGen-Bench, a 2,700-document Bengali benchmark with human-written, English-to-Bengali machine-translated and LLM-generated classes, and a stylometric detector that needs no Bengali-specific tooling. Translatedness did not transfer between the two NMT systems we tested: a detector trained only on NLLB-200 output recognises mBART-50 output 5.67% of the time, and the reverse 0.00%, while mixed-system training reaches 99.3% on held-out documents of a seen system. A detector trained on dedicated NMT catches an LLM prompted to translate 20 times in 100, against 97 when the same model generates. Across four snapshots (2018–2026), the detector-flagged non-human share rises from 2.13% to 6.20%, and after 2023 the flagged movement is on the translation side: the MT-flagged share climbs from 0.30% to 3.48%, while the LLM-flagged share peaks and falls. These are flagged shares, not validated prevalence: on the only independent Bengali corpus available, which has no translated text, the detector scores **62.7%** against 97.3% in-benchmark. The benchmark, checkpoints and scripts will be released.

1 Introduction

Machine translation for Asian languages is built on monolingual web text. Back-translation converts target-side monolingual data into synthetic parallel data (Edunov et al., 2018); bitext mining scores candidate pairs drawn from crawls; and the pre-trained models that initialise low-resource systems are trained on Common Crawl derivatives such as CCNet (Wenzek et al., 2020), OSCAR (Abadji et al., 2022) and MADLAD-400 (Kudugunta et al., 2023). For Bengali and its Indic neighbours, where curated parallel data is scarce, the dependence is heavier still: Samanantar (Ramesh et al., 2022) and IndicTrans2 (Gala et al., 2023) both lean on web-derived material, as does NLLB-200 (NLLB Team et al., 2024).

That pipeline closes on itself if the web is already machine-translated. Thompson et al. (2024) show, using multi-way parallelism, that a large share of web text is MT output and that the effect is worst exactly where data is scarcest. Training on such data is a form of self-consumption (Shumailov et al., 2024), and the translationese it carries is known to distort both models and the test sets used to evaluate them (Riley et al., 2020; Zhang and Toral, 2019; Bogoychev and Sennrich, 2019). Audits of web-crawled multilingual corpora have documented serious quality problems in low-resource languages (Kreutzer et al., 2022; Caswell et al., 2020), but none measures the machine-translated share of a single Asian language across the period in which neural MT and instruction-tuned LLMs both reached scale.

Bengali is an instructive case. It has more than 270 million speakers and an active web presence, it is a target language in every major Indic MT effort, and yet no resource separates human-written, machine-translated and directly LLM-generated Bengali text. Without that separation one cannot tell whether a rise in machine-produced web text is an MT problem or a generation problem – and for a translation workshop those are different problems with different remedies. The gap is instrumentation: before proposing filters for MT training corpora, we need to be able to measure what is in them.

Research questions. **RQ1** (Sections 3–4): can human-written, machine-translated and LLM-generated Bengali be told apart when provenance is known? We introduce **BanglaGen-Bench**, a labelled three-class benchmark with a disclosed multi-system mix per synthetic class, to be released

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 104–118
November 9, 2026 ©2026 Association for Computational Linguistics

with checkpoints and code. **RQ2** (Section 5): does learned translatedness transfer across MT systems, and across one model's roles as generator and as translator? Here every label is known, and transfer fails twice. **RQ3** (Section 6): how do detector-flagged shares of Bengali Common Crawl change from 2018 to 2026, and do they move toward translation or generation? RQ3 concerns predicted labels, not verified provenance; Section 7 reports the external check that bounds it. None of these questions is new on its own (Section 2); what is new is asking them together, for one low-resource Asian language, with translation and generation kept apart.

What this paper measures, and what it does not. No oracle exists for whether a given web document was machine-produced. Our detector is trained on *constructed* classes – text we translated and text we generated – and applied to text whose provenance nobody knows, so everything downstream estimates the share of documents *a particular detector flags*, not prevalence; we use *detector-flagged share* in that literal sense. The gap is real: on the only independently labelled Bengali corpus we could obtain, the detector scores 62.7% against 97.3% in-benchmark, and that corpus has no translated text (Section 7), so we call the trend preliminary. The cross-system results of Section 5 do not share this limit, because there we produced every document.

2 Related Work

Translationese and its detection. That translated text is systematically distinguishable from originally authored text is a long-standing result (Baroni and Bernardini, 2006; Koppel and Ordan, 2011; Rabinovich and Wintner, 2015), established mostly on clean, parallel-aligned corpora in high-resource European languages. Our setting differs on three axes at once: noisy extracted web text, a low-resource Asian language with no comparable resources, and a three-way taxonomy that separates translation from direct generation. The last matters because a two-class human-versus-AI scheme cannot see the flagged composition shift we report in Section 6.2.

Machine-translated text in MT training data. Translationese in synthetic data measurably changes what a system learns (Edunov et al., 2018; Bogoychev and Sennrich, 2019) and what evaluation measures (Zhang and Toral, 2019); tagging it explicitly helps (Riley et al., 2020). Thompson et al. (2024) is the closest prior work to our measurement: they detect MT output on the web through multi-way parallelism across languages, and find low-resource languages worst affected. Our approach is complementary and independent in its failure modes – monolingual stylometry within one language, tracked over time, and separating MT from LLM generation, neither of which multi-way parallelism addresses. Marie et al. (2021) is the methodological precedent we try to follow in reporting what failed.

Crawl quality for low-resource languages. Kreutzer et al. (2022) audit web-crawled multilingual datasets and find substantial mislabelling and low-quality content, concentrated in exactly the languages that most need the data; Caswell et al. (2020) document the language-ID failures underneath. These audits are cross-sectional. He et al. (2026) find off-the-shelf detectors weak on LLM-dominant crawl and web data, which is one motivation for our multi-method audit; Longpre et al. (2024) study data age, coverage and quality; Cheng et al. (2025) survey contamination detection more broadly.

Indic and Bengali resources. IndicCorp (Kakwani et al., 2020), Samanantar (Ramesh et al., 2022), IndicTrans2 (Gala et al., 2023), MADLAD-400 (Kudugunta et al., 2023) and BanglaBERT (Bhattacharjee et al., 2022) filter for language, duplicates and quality, but none filters for, or reports, machine-produced source text. That is the practical gap this paper addresses.

Machine-generated text detection. DetectGPT (Mitchell et al., 2023) and Binoculars (Hans et al., 2024) use probability curvature and cross-model perplexity; both need a strong target-language reference model, which is harder to obtain for Bengali, whereas our stylometric method needs none. Wu et al. (2025) identify out-of-distribution generalisation and weak evaluation as the open problems in this area, and our Section 7 is an instance of both. MULTITuDE (Macko et al., 2023) and RAID (Dugan et al., 2024) cover multilingual and adversarial robustness but not Bengali. Ansary (2025), Islam et al. (2025) and Ammar et al. (2025) report 91–99% on curated two-class Bengali and Urdu detection; none separates translation from generation or tests noisy crawl text. Dycke et al. (2026)

warn that narrow AI classes fail on messier real AI text, a limitation our classes share and our external evaluation confirms.

3 BanglaGen-Bench

All three classes are built from, or matched to, the same pool of Bengali Wikipedia article titles, so that provenance rather than topic is the primary axis of variation. Table 1 summarises the composition.

Human-written class. Human text comes from two sources with strong prior odds against machine authorship: a stratified, fastText-filtered sample of `CC-MAIN-2018-17`, which predates instruction-tuned LLMs at web scale, and randomly drawn current Bengali Wikipedia articles, whose editorial norms make wholesale LLM authorship comparatively rare (576 and 324 documents respectively). Neither is assumed to be free of machine content, and Section 6.3 treats that assumption as the load-bearing one it is.

Machine-translated class. The mt class was produced by translating extracts of English Wikipedia articles into Bengali (en $\rightarrow$ bn), matched to the shared title pool through interlanguage links, initially with NLLB-200 distilled 600M (NLLB Team et al., 2024) alone. Each mt document is therefore a single translation step from an English source, the same kind of operation that puts machine-translated Bengali on the web; no round trip through a pivot language is involved. That configuration failed, in a way that turned out to be the most transferable result in the paper; Section 5 reports it. The final class mixes three systems, each translating English extracts from topic-disjoint title sets.

LLM-generated class. The llm class was directly prompted, which separates generation from translation at the point of construction. The disclosed mix uses Llama 3.1 8B (Llama Team, AI @ Meta et al., 2024) and Qwen2.5 7B (Qwen Team, 2024), yielding 308 Llama and 592 Qwen documents. Llama only partly fit the available 4GB GPU while Qwen2.5 fit and ran roughly 3.4$\times$ faster, which produced the uneven split; the constraint is disclosed rather than presented as a design choice.

Crawl samples and corpus integrity. Four snapshots run through an identical fetch, clean and dedup pipeline: `CC-MAIN-2018-17` as a pre-LLM baseline, `CC-MAIN-2023-06` as an early post-ChatGPT point, `CC-MAIN-2024-30` as an intermediate one and `CC-MAIN-2026-30` as the current one. Two integrity passes collapse within-snapshot near-duplicates (Jaccard 0.5; a government-portal boilerplate family disproportionately affected 2018) and exclude evaluation documents matching or near-duplicating benchmark construction data. Final sizes are 6,579 / 9,752 / 12,281 / 9,823 documents over 2,453 / 3,993 / 4,793 / 4,435 domains. Domain, not document, is the statistical unit throughout (Section 6.4).

Class	Source system / origin	n
human	`CC-MAIN-2018-17` sample	576
	Bengali Wikipedia	324
mt	NLLB-200 distilled 600M	626
	mBART-50	204
	Mistral 7B (training only)	70
llm	Qwen2.5 7B	592
	Llama 3.1 8B	308
Total		2,700

Table 1: Benchmark composition. All mt documents are extracts of English Wikipedia articles translated into Bengali in a single step (en $\rightarrow$ bn). The multi-system mix is a consequence of Section 5, not a design preference: a single-system mt class does not generalise. The 70 Mistral translations were added after the unseen-system test of Section 5.2 and are training-only, so no held-out-system claim is made for Mistral in the deployed detector.

4 Detector

Stylometric baseline. The deployed detector is a linear SVM over word-boundary-aware character n-grams ($n \in \{2, 3, 4\}$), TF-IDF weighted. It is deliberately language-agnostic: it requires no Bengali morphological tooling and no reference LLM at inference, which is what makes it portable to other low-resource Asian languages. It reaches 97.28% test accuracy on the 405-document test split and 95.5%$\pm$0.7% under five-fold cross-validation. Class balancing left 600 already-collected documents (200 per class) unused; the deployed model scores 96.67% on that disjoint holdout, with per-class recall within 1.5 points of the primary test, so the confusion matrix is not an artifact of one split.

A quality-estimation signal, explored and demoted. Following Wastl et al. (2025) and unsupervised quality estimation (Fomicheva et al., 2020), we also tried an MT-native feature: the NLLB-200 log-probability of back-translating each

document's leading span to English. It adds 0.99 test points, but McNemar's exact test finds the gain not significant ($p = 0.21875$), so it is used for no result here (Appendix A.1).

Calibration. Test accuracy is near 100% in the top nine decision-margin deciles and 78.0% (n=41) in the lowest; Platt-scaled Brier 0.022 against a constant baseline's 0.026.

Cross-architecture checks. Two transformer encoders were fine-tuned on the identical split and applied to the same snapshots under the same leakage guard. BanglaBERT (Bhattacharjee et al., 2022) is Bengali-pretrained and reaches 98.02% test accuracy. XLM-R-base (Conneau et al., 2020) is massively multilingual rather than Bengali-specialised, fine-tuned by LoRA (Hu et al., 2022) with rank-8 adapters (0.32% of parameters trainable), because full fine-tuning exceeded a 4GB GPU.

Why no off-the-shelf baseline. Zero-shot detectors such as DetectGPT and Binoculars need a strong Bengali reference model (Section 2), and published Bengali detectors are two-class, so none can score the mt class without retraining on our classes; the two encoders above are that retrained, like-for-like comparison.

5 Translatedness Did Not Transfer Across NMT Systems

This section reports the results whose provenance labels we control completely, and which therefore do not inherit the epistemic limit of Section 1. Their scope is narrow – two NMT systems, one source language, one domain – so non-invariance is an existence result: it shows that invariance cannot be assumed, not how often it fails.

5.1 Single-system training learns a system

We built the mt class with NLLB-200 alone, as most work of this kind would. Evaluated against held-out mBART-50 (Tang et al., 2021) translations of the same kind of source material, that detector recognised 5.67% of them. Trained the other way round – mBART-50 only, evaluated on NLLB-200 – it recognised none at all (Table 2). Adding 300 topic-disjoint mBART documents to training raised accuracy on 300 held-out mBART documents to 99.3% (298/300): held-out documents of a system seen in training, not an unseen system. (This experiment used a larger mBART-50 pool than the 204 documents in Table 1.)

Training configuration	Evaluated on	Detected
mt class, single system		
NLLB-200 only	mBART-50	5.67%
mBART-50 only	NLLB-200	0.00%
mt class, mixed systems		
NLLB + mBART	mBART-50†	99.3%
llm class, single system		
Llama only	Qwen2.5	97.97%
Qwen2.5 only	Llama	99.35%
a genuinely unseen system, two roles		
llm class, Mistral unseen	Mistral, generating	97/100
mt class, Mistral unseen	Mistral, translating	20/100‡

Table 2: Cross-system generalisation. †Held-out *documents* from a system present in training, not a held-out system; the distinction is load-bearing and we do not elide it. ‡Measured before the 70 Mistral translations of Table 1 entered training. The last two rows are the same unseen model in two roles: generation transfers, translation does not.

The magnitude matters less than the asymmetry with the llm class, where single-system training transfers almost perfectly (97.97% and 99.35% across systems). Whatever the detector learns from one NMT system's output is close to disjoint from what it needs for another's, while whatever it learns from one instruction-tuned LLM largely covers the next. The natural reading, for the two systems we tested, is that "machine-translated" as an empirically learnable category is, at single-system scale, a set of system-specific artifacts rather than a property of the translation process. Detectors, filters and quality-estimation heuristics built on one NMT system's output should be assumed not to hold for another until tested, and corpus-cleaning pipelines that report an MT-filtering step should report which systems that step was calibrated on.

5.2 Dedicated NMT and LLMs-as-translators are different problems

The last two rows of Table 2 are the result with the shortest shelf-life and the largest consequences. Mistral 7B, held out from the llm class entirely, is caught 97 times in 100 when it *generates* Bengali. The same model, prompted to translate fresh English Wikipedia extracts and absent from the mt class at the time, is caught 20 times in 100 when it *translates* into Bengali; half of its translations are labelled llm and 30 of 100 human. One unseen model, two roles, a nearly five-fold difference in detectability.

The mt-class repair of Section 5.1 holds for the

dedicated NMT systems it was trained on and does not extend to an LLM prompted to translate. We then added 70 topic-disjoint Mistral translations to the mt class (Table 1), which repairs this one model only. Mistral thus plays four roles, which we keep apart: unseen generator (97/100), unseen translator before the repair (20/100), source of 70 training translations after it, and one of the paraphrasers of Section 5.3. Only the first two are unseen-system results, and the deployed detector has not been tested on a dedicated NMT system absent from its training data. As prompting displaces dedicated NMT, this blind spot widens without anything in the detector changing, so any estimate of the machine-translated share of web text, ours included, understates LLM-produced translation; the TPR sweep of Section 6.3 treats it as a parameter (Appendix A.3).

5.3 Paraphrase robustness, and what one paraphraser revealed

The standard evasion against text-provenance detectors is paraphrase (Krishna et al., 2023; Hu et al., 2023; Zha et al., 2025). We paraphrased 80 held-out mt- and llm-class documents with three local models (Llama 3.1 8B, Qwen2.5 7B, Mistral 7B) and, on a 16-document subsample, with Claude Sonnet 5, then re-scored everything.

The result replicates the system-specificity of Section 5.1 on the attack side. **Mistral evaded the detector 63.75% of the time**; Llama and Qwen evaded it 0% and 1.3% (n=77–78), so a two-paraphraser evaluation would have concluded the detector was robust. Train-split augmentation with 60 extra paraphrases across the three local models closes Mistral evasion to 0.00%, and expanding the attack to n=247 and n=251 holds Llama and Qwen evasion at 0.40%, tightening one-sided Clopper–Pearson upper bounds to 1.91% and 1.88%. Claude Sonnet 5 was never in the augmentation and is the only paraphraser genuinely held out from selection: evasion falls from 18.75% to 12.5% (n=16), a one-document movement from which we take the direction and not the magnitude.

The repair has a cost, and it runs along the MT/LLM axis again. Generalisation to an unseen LLM generator regresses from 97% to 78%, while mt-side detection of Mistral-translated text moves from 20% to 96% – though Mistral is not fully unseen there, since 70 mt-class training documents are Mistral translations (Table 1), none overlapping this test. Both detector variants will be released: a robustness-for-generalisation trade, not a strict improvement.

Snap.	n	Raw	ACC	Cluster 95% CI
2018	6,579	2.13%	1.50%	[0.21%, 2.32%]
2023	9,752	5.65%	3.63%	[2.30%, 5.72%]
2024	12,281	5.14%	3.11%	[1.74%, 5.26%]
2026	9,823	6.20%	4.20%	[1.92%, 6.31%]

Table 3: Detector-flagged non-human share per snapshot. ACC is the 2018-anchored Adjusted Classify & Count estimate, whose estimand Section 6.3 qualifies; the rightmost column is its 95% interval under a bootstrap that resamples domains rather than documents.

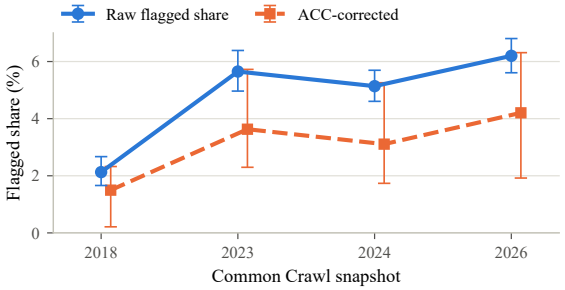


Figure 1: Detector-flagged share by snapshot with cluster-robust 95% intervals, plotting Table 3. The 2018-to-2023 step and the 2023–2026 plateau are the two features every check in this paper tests.

6 Detector-Flagged Shares in Bengali Common Crawl

We now apply the deployed detector to each snapshot and define the **detector-flagged share** as the predicted non-human (mt plus llm) proportion. Everything in this section is subject to the caveat of Section 1 and quantified in Section 7: it measures changes in predicted labels, not verified provenance, and a rising flagged share may reflect change on the web, detector errors that move with the snapshots, or both.

6.1 Rates

The raw rate rises 2.13% → 5.65% → 5.14% → 6.20%, a 2.9× relative increase from 2018 to 2026, or 2.8× under ACC. Every comparison against 2018 is significant (Holm-corrected two-proportion z, $p_{\text{holm}} \approx 0$). Neither 2023-vs-2024 nor 2023-vs-2026 reaches significance ($p_{\text{holm}} = 0.188$); 2024-vs-2026 does ($p_{\text{holm}} = 0.002$), a small naive-rate rise. Cohen's h against 2018 is small ($h \leq 0.21$), as it must be for low proportions: a large relative change in a small absolute share.

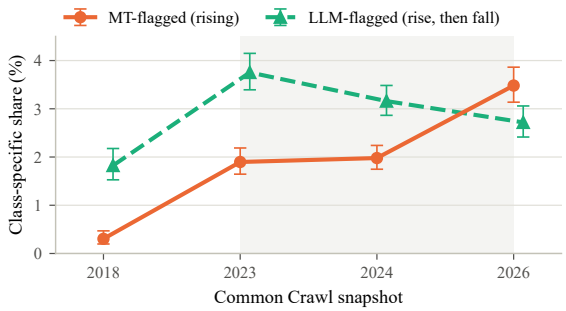


Figure 2: mt-flagged against llm-flagged share (Wilson 95% CI). The shaded 2023–2026 span is where the two move in opposite directions, which the combined rate of Figure 1 hides. Only the mt-side rise survives the matched-length retrain of Appendix A.7.

The plateau is not a power artifact. For the tightest post-2023 pair, 2024-vs-2026, a two-proportion power calculation puts the smallest step detectable at 80% power under the Holm worst-case threshold at 0.83 percentage points, and no observed post-2023 gap approaches it.

6.2 The composition shift the combined rate hides

Tested apart, the two flagged shares move in opposite directions, and for this venue that is the result.

The mt-flagged share rises monotonically (0.30% → 1.90% → 1.98% → 3.48%), is significant against 2018 throughout ($p_{\text{holm}} \approx 0$), is flat across 2023-vs-2024 ($p_{\text{holm}} = 0.663$), and rises in both 2023-vs-2026 and 2024-vs-2026 ($p_{\text{holm}} < 10^{-10}$). The llm-flagged share rises and then falls (1.82% → 3.75% → 3.16% → 2.72%), also significant against 2018, with declines across 2023-vs-2024 ($p_{\text{holm}} = 0.032$) and 2023-vs-2026 ($p_{\text{holm}} < 10^{-3}$) but not 2024-vs-2026 ($p_{\text{holm}} = 0.055$).

The 2024-to-2026 rise in the combined flagged share therefore comes from the mt side. A two-class human-versus-AI scheme would miss this inversion entirely. It also survives the strongest internal challenge we could mount: under the matched-length retrain that replaces the llm class entirely with length-matched documents (Appendix A.7), the mt-side rise persists across seeds while the llm-side movement does not. Read conservatively, we can defend the mt-side rise in what the detector flags, not the llm-side trajectory; whether the web itself holds more machine-translated Bengali, no external corpus available to us can say (Section 7).

6.3 Calibration, and an estimand we do not claim

Classify-and-count is biased under class imbalance whenever per-class error rates differ, which is our situation, so we apply Adjusted Classify & Count (Forman, 2008), inverting the confusion matrix and clipping infeasible negative solutions to zero.

The synthetic test-set false-positive rate is 4.48%. The 2018 snapshot predates both ChatGPT and GPT-2, so it is all but certainly free of LLM-generated text by temporal necessity; it nonetheless flags 2.13% of its documents as non-human (140 of 6,579; 95% CI [1.79%, 2.51%]). The correction uses the same quantity measured on a vintage-retrained detector (Appendix A.6), 2.10% (138 of 6,579); the FPR sweep (Figure 3) spans both values and no conclusion changes between them. The correction rescales the human row while preserving the llm:mt misclassification ratio, giving 1.50% / 3.63% / 3.11% / 4.20%; the synthetic-FPR alternative gives 1.13% / 3.02% / 2.43% / 1.99%. With the 2018 anchor every post-2018 year is significant against 2018 ($p_{\text{holm}} \leq 0.004$); with the synthetic FPR only 2023 is. The later years remain mutually indistinguishable ($p_{\text{holm}} \geq 0.09$, rising to ≥ 0.58 under cluster-robust testing). Across 31 swept FPR values from 2% to 5%, 2018 stays below all later years with a minimum 0.86-point gap (Figure 3), well powered throughout (Appendix A.2), though the later years' relative order flips across the sweep, which reinforces that they are not separable.

The anchor is class-asymmetric, and we do not pretend otherwise. Chronology rules out LLM generation in 2018. It does not rule out machine translation. Statistical MT, templated automation and text-spinning all predate 2018 and would have been on the Bengali web then. The two components of the 2.13% are therefore not equally well identified. Any 2018 document flagged `llm` must be a false positive, so that component is a direct measurement of the human-to-llm error rate. A 2018 document flagged `mt` may be a correct detection of pre-neural machine translation, so that component is only an upper bound on the human-to-mt error rate, and treating all of it as error subtracts real signal along with noise.

Two consequences follow. First, the combined ACC figure has a mixed estimand and should be read as a bound rather than a point estimate; using an FPR that is too high depresses every corrected rate, so the levels in Table 3 are conservative, and

the FPR sweep shows the ordering is insensitive to the choice anywhere in 2–5%. Second, the 2026 end of the mt-side trend is understated for an unrelated reason: LLM-produced translation was detected at 20 in 100 before the targeted Mistral repair (Section 5.2). A properly identified estimate would calibrate the two classes separately against Bengali text of known provenance; no such corpus was available to us, and we do not claim that number.

Bounding the sensitivity side. No real-text oracle grounds the true-positive rate either, so we bound rather than assume it. The llm-class synthetic TPR of 100% updates to the third-party estimate of 97% directly. The mt-class synthetic TPR of roughly 96% diverges sharply from the 20% measured on Mistral-translated text, but that 20% is the mechanism-specific worst case of Section 5.2 rather than a representative real-world rate, so we sweep the mixing share $w \in [0, 1]$ of real-world MT behaving like the worst case. The finding holds across the large majority of the range and crosses zero only near $w \approx 0.85$, which would require more than 85% of real Bengali web MT to come from prompting a general-purpose LLM rather than from cheaper, faster dedicated NMT (Appendix A.3).

6.4 What survives

Web documents from the same domain are not independent, so document-level intervals overstate precision. We resample each snapshot's *domain list* over 5,000 iterations with a shared per-iteration confusion-matrix draw (Appendix A.4). Naive intervals widen by 26–55%; ACC-corrected intervals by only 2–5%, because confusion-matrix resampling noise from the 405-document test set already dominates their width. Every significance conclusion survives: 2018 stays below all later years ($p_{\text{holm}} < 0.004$), and among the later years only 2024-vs-2026 separates, and only on naive rates: under cluster-robust ACC it does not ($p_{\text{holm}} = 0.58$).

Both alternative architectures place 2018 significantly below every later year and agree on the shape – flat 2023–2024, higher 2026 naive rates – while disagreeing substantially on level (BanglaBERT 1.08% / 4.40% / 4.01% / 4.90%; XLM-R-LoRA 7.39% / 11.25% / 10.56% / 12.65%). We therefore treat no architecture's absolute level as authoritative and read only the shape. Per-document agreement on the three-class label is only fair by Cohen's κ (0.20–0.28 against BanglaBERT, with 92–97% identical labels; 0.13–0.24 against XLM-R-LoRA, with 86–92%), which is what κ does at this prevalence: chance agreement alone is 82–97% and unequal marginals cap the attainable value well below unity (Appendix A.5). This is mutual consistency, not independent validation.

Two independent retrains on generators unseen during the original work – Claude Sonnet 5 (24 training, 16 held out) and Llama 3.2 3B (79 training, 60 held out) – reproduce the original rates to within 0.07 and 0.29 points with the identical significance pattern (Appendix A.6): the detector is not of the wrong vintage, though the classes remain constructed.

Corpus-quality metrics (type-token ratio, duplicate 5-grams, a compression proxy (Shaib et al., 2025)) move toward lower diversity, and a classifier two-sample test (Lopez-Paz and Oquab, 2017) separates our human class from the live snapshots increasingly well (AUC 0.68 → 0.93) – equally consistent with our training data drifting out of distribution. Movement is established; its cause is not.

No method applied to 2018-vs-2026 reverses its direction (Appendix A.8); only the llm-only per-class rate fails the matched-length retrain.

A regional split. By top-level domain, Bangladesh-hosted (`.bd`) content is flagged more often than India-hosted (`.in`) content in every testable year (11.11% against 4.19% in 2026; $p_{\text{holm}} < 0.001$ in 2023, 2024 and 2026). Our human class is drawn mostly from `.bd` sources, so we trust this only because calibration, leave-one-domain-out and organisation-level checks all hold (Appendix A.11); it covers only the 10.4% of documents with a country TLD (Limitations).

A finding that failed its own stress test. A flagged-share spike in the ChatGPT launch window survived class-composition and top-k domain checks but dissolved under inverse-propensity domain reweighting (Appendix A.12): single-crawl estimates can manufacture temporal spikes.

7 The External Check, and What It Costs the Trend

Every result in Section 6 tests internal consistency; none tests the classes against independently sourced labels. This section does, and its outcome is the most important caveat in the paper.

Genre	n/class	Human	AI
Newspaper Article	159	98.7%	38.4%
Scientific Abstract	81	18.5%	81.5%
Email	27	33.3%	63.0%
Poem	15	60.0%	100.0%
Short Story	15	60.0%	93.3%
Office Order	3	33.3%	100.0%

Table 4: Per-genre accuracy on the independently labelled dataset of Ansary (2025). *Human* is the share of human-labelled documents predicted human; *AI* the share of AI-labelled documents predicted mt or llm. Office Order, at three documents per class, is too small to weigh alone.

Among external Bengali provenance corpora we could locate, a 40,000-document Zenodo corpus is embargoed until December 2026 and the rest are paywalled or lack a peer-reviewed description. The dataset of Ansary (2025) is public: 600 documents, 300 human and 300 ChatGPT-generated, matched by genre, independently labelled and tied to a peer-reviewed paper. Its labels are binary where our detector predicts three classes and it contains no translation content, so it speaks only to the human-versus-llm axis. Taking the most permissive fair reading of correctness for an AI-labelled document – predicted mt *or* llm – human recall is 66.7% (200/300), AI recall 58.7% (176/300) and binary accuracy **62.7%** (376/600), against 97.28% in-benchmark.

The two failure modes are mechanistically distinct (Table 4). Human-authored Scientific Abstracts carry a dense register of English-loanword technical terms transliterated into Bengali script, entirely absent from our human class's Wikipedia and CC-2018 distribution, which plausibly explains why genuine human text in that genre is so often called non-human. For Newspaper Article – the genre closest both to our training register and to real web content – the low 38.4% AI-catch rate is not our length confound working in the generator's favour: the ChatGPT newspaper text is *shorter* on average than the paired human text (637 against 1,586 characters), which our documented length sensitivity should make easier to catch. It is a style-generalisation gap against an unseen generator. Of the AI-labelled documents caught, 161 of 176 (91.5%) are predicted llm rather than mt, as expected without translation.

What the check cannot reach. The corpus contains no machine-translated text, so it tests only the human-versus-non-human axis under distribution shift and gives no external support to the translation side of Section 6.2. We know of no independently labelled corpus of machine-translated Bengali web text; building one is the most useful next step.

What this permits and forbids. It forbids reading Table 3 as prevalence: internal agreement cannot repair a 62.7% transfer rate, since the internal checks share the construction bias that causes it. It permits the weaker claim we make: a fixed detector, held constant across four identically processed snapshots, flags a rising share over time, and the flagged rise decomposes toward the mt class. It does not establish a translation-driven shift in the true composition of the web. A detector can be badly calibrated in level and still informative in difference, provided its bias does not move with the snapshots – an assumption the C2ST result of Section 6.4 partially undermines. Hence we call the trend preliminary.

8 Implications for Asian-Language MT Resources

Three consequences follow for anyone building MT resources from Asian web text. **MT filters must disclose their calibration systems**: a detector trained on one NMT system's output recognised another's at 5.67% and 0.00% (Section 5.1), so pipelines that filter machine-translated text should name the systems the filter was calibrated on, as they already state language-ID thresholds. **LLM translation is the uncovered case**: filters calibrated on dedicated NMT miss it (Section 5.2). **The monolingual side may be moving**: if even part of the rise in the mt-flagged share, from 0.30% to 3.48%, is real, target-side data for Bengali back-translation holds a growing share of MT output – the self-consumption setting of Shumailov et al. (2024), to which synthetic-data quality is sensitive (Bogoychev and Sennrich, 2019) – and none of the corpus pipelines of Section 2 measures it.

Limitations

No independently sourced ground truth for real web text. This is the central epistemic limit, and it is stated in the abstract, in Section 1 and in Section 7 rather than here alone. All detectors in this paper train on constructed classes; the external check gives 62.7% against 97.28% in-benchmark, and it covers only the human-versus-non-human axis: with no translated text in it, the mt class and

the translation side of the trend have no external validation. Vintage retrains address the objection that the detector is out of date. They do not address the objection that the classes are constructed.

A held-out-system gap in both directions. Single-system MT training fails in both directions (0.00% and 5.67%), unlike single-system LLM training (97.97%, 99.35%), and that non-invariance rests on two NMT systems. The 99.3% mixed-training result is held-out documents, not a held-out system, and an unseen further NMT system may fail similarly. The mt-side repair holds for the dedicated NMT systems it was trained on, not for an LLM prompted to translate; the later Mistral repair covers one model only, and the deployed detector has not been tested on a dedicated NMT system absent from its training data.

What the mt class models, and what it may confound. Every mt document is a single en → bn translation of an English Wikipedia extract. Machine-translated Bengali on the web may also start from source languages other than English, from registers other than encyclopaedic prose, and from systems we did not use, commercial MT among them; the system-specificity of Section 5 is shown for one source language, one domain and two NMT systems. Title matching aligns topic across classes but not content origin: mt documents carry content first written in English, while human documents are native Bengali, so part of what separates mt from human may be English-origin content rather than translation artifacts.

Evaluations not reported here. We report the three-class model throughout; the external check of Section 7 reads it as two-class by collapsing its mt and llm predictions, and a directly trained human-versus-non-human detector is not reported. We also do not report how source titles are distributed across training, development and test splits; a title-grouped re-split would rule out topical leakage between them. Section 4 explains why no off-the-shelf detector is compared.

A length confound, partly resolved. The llm class has a fixed generation cap and is shorter than the others; a length-only classifier reaches 69.4%±1.8%, above chance, so length contributes without dominating. Full-class length parity was measured to be infeasible on the available hardware, so we regenerated 100 documents at a median 7.2× the original length; 97 of 100 are still classified llm, with no reciprocal false-positive rise among human and mt documents in that range (97.0% and 99.1% still correct). This supports inference-time robustness and does not remove the training confound. Crawl document length itself rises from 2018 to 2026, which should bias the measured trend downward (Wang et al., 2025).

The regional finding covers a minority of the corpus. Clear country TLDs cover 10.4% of the pooled corpus (8.5–15.9% per year). Targeted index-guided collection grew every `.in` cell (79→81; 432→448; 653→682; 406→423) while changing the rate by at most 0.068 points, and 2018 gained two documents from 25 independently confirmed segments, which indicates a structural ccTLD ceiling rather than a sampling failure (Appendix A.9). Proxy geolocation would add bias: hosting location is not authorship under a CDN, and language is not country. ccTLD documents are shorter (3,560 against 4,534 characters), carry lower human-classification margins (1.76 against 1.99) and are far more government-heavy (33.1% against 3.4%), as expected when `.gov.bd` and `.gov.in` structurally require a country TLD. The finding applies to that subpopulation.

Small and partly circular robustness samples. The paraphrase battery cannot exclude true evasion near 6% for models showing none at n=80, though the expansion to $n \approx 250$ tightens this to under 2%. Augmentation and attack data share the three local models; only Claude Sonnet 5 is genuinely held out from selection, at n=16. Extraction noise leaves human predictions nearly unchanged (0.0–0.2% flips at 5% character noise) but makes mt and llm predictions fragile (22.2% and 20.0%), which is consistent with the thinner minority-class decision margins found in Section 4.

Scale and sampling. Four time points give a coarse rather than continuous trend, and 900 benchmark documents per class from two or three systems each is small. Language ID is stable across snapshots (mean confidence 0.984–0.989, 2.4–3.4% borderline), and restricting the trend to the roughly 90 domains common to all four snapshots preserves the shape while losing power, which inverse-propensity weighting recovers at the same conclusion (Appendix A.10).

Ethics Statement

All text comes from public web sources (Common Crawl) and from Bengali and English Wikipedia, consistent with their terms of use for research. We do not target, collect or release personal or identifying information.

We are conscious of the dual-use potential of provenance detection for a low-resource language. A detector flagging likely machine-translated or LLM-generated Bengali could be repurposed to suppress or deprioritise non-native speakers, people using translation and writing tools as an accessibility aid, or Bengali communities whose style diverges from our training distribution. The external evaluation in Section 7 makes the risk concrete rather than theoretical: 81.5% of genuine human-authored Bengali scientific abstracts in an independent corpus were flagged as non-human by our detector, because their register is absent from our human class. We will release BanglaGen-Bench and the detector as research artifacts for corpus analysis and filtering research, and we caution explicitly against deployment in content moderation or individual-authorship judgements, given AI-detector false-positive risks and the specific risk of penalising legitimate non-native or assistive use of translation and writing tools.

Acknowledgments

We thank the three anonymous reviewers, whose comments shaped the research questions of Section 1, the scope statements of Sections 5 and 7, and the Limitations. We also thank Farig Yousuf Sadeque (BRAC University) for his contributions to this work.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Muhammad Ammar, Hadiya Murad Hadi, and Usman Majeed Butt. 2025. AI-generated text detection in low-resource languages: A case study on Urdu. *arXiv preprint arXiv:2510.16573*.

Md. Siam Ansary. 2025. A novel dataset for identifying AI generated Bengali texts. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A survey on data contamination for large language models. *arXiv preprint arXiv:2502.14425*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2405.07940.

Nils Dycke, Marina Sakharova, Nico Daheim, and Iryna Gurevych. 2026. "Your AI Text is not Mine": Redefining and evaluating AI-generated text detection under realistic assumptions. *arXiv preprint arXiv:2606.04906*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled languages of India. *Transactions on Machine Learning Research*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Rounak Saha, Yuxin Wen, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Sichang Steven He, Calvin Ardi, Ramesh Govindan, and Harsha V. Madhyastha. 2026. DeGenTWeb: A first look at LLM-dominant websites. *arXiv preprint arXiv:2605.00087*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2307.03838.

Md. Rakibul Islam, Most. Sharmin Sultana Samu, Md. Zahid Hossain, Farhad Uz Zaman, and Md. Kamrozzaman Bhuiyan. 2025. Detecting AI-generated paraphrases in Bengali: A comparative study of zero-shot and fine-tuned transformers. *Preprint*, arXiv:2512.21709.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2303.13408.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Llama Team, AI @ Meta, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

David Lopez-Paz and Maxime Oquab. 2017. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

Jean Maillard, et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C. Wallace. 2025. Measuring AI "Slop" in text. *arXiv preprint arXiv:2509.19163*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Yitong Wang, Zhongping Zhang, Margherita Piana, Zheng Zhou, Peter Gerstoft, and Bryan A. Plummer. 2025. Real, fake, or manipulated? detecting machine-influenced text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15022–15037, Suzhou, China. Association for Computational Linguistics.

Michelle Wastl, Jannis Vamvas, and Rico Sennrich. 2025. Machine translation models are zero-shot detectors of translation direction. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Yiwei Zha, Rui Min, and Shanu Sushmita. 2025. PADBen: A comprehensive benchmark for evaluating AI text detectors against paraphrase attacks. ArXiv:2511.00416.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Supplementary Material

Full derivations and setups for results summarised in the main text. Nothing here changes a conclusion stated in the body.

A.1 Hybrid-signal significance testing

The hybrid concatenates the mean per-token NLLB-200 back-translation log-probability to the n-gram matrix as a translation-confidence feature (Table 5). Its gain over the deployed stylometric model is four net documents on the 405-document test set, concentrated on predicted MT-versus-LLM confusion. McNemar's exact test on the MT-versus-LLM cell involves two discordant pairs ($p = 0.5$); on the full test set, six ($p = 0.21875$). Powering the class-specific test would require an order of magnitude more discordant pairs than this split yields. Rather than bill an unproven effect as a contribution, we report the signal as explored but not adopted.

Model	Dev acc.	Test acc.
Stylometric-only (deployed)	95.80%	97.28%
+ transl.-conf. (explored)	96.30%	98.27%

Table 5: Detector ablation on the 2,700-document benchmark. The stylometric-only model is the one deployed for every result in this paper; the hybrid is reported and demoted, not used.

A.2 Minimum detectable effect across the FPR sweep

The 0.86-point minimum gap of Section 6.3 describes 2018-vs-2026, whereas the 0.83-point floor of Section 6.1 describes 2024-vs-2026; the two are not interchangeable. Recomputing the minimum detectable effect specifically for 2018-vs-2026 at each swept FPR ($n = 6{,}579$ and $n = 9{,}823$) gives a floor of 0.58–0.69 points across the sweep, so the observed gap exceeds its own matched floor by a factor of 1.2–4.0 at every point. Script: `scripts/compute_minimum_detectable_effect.py`.

A.3 TPR-sensitivity sweep

The true-positive rate has no equivalent of the FPR's chronological anchor: it comes entirely from the synthetic test set, and no oracle exists for "this document is definitely machine-produced". We therefore bound it. The llm-class synthetic TPR of 100% is close to the third-party estimate of 97% on Mistral-generated text, so we apply that update directly. The mt-class synthetic TPR of roughly 96% diverges sharply from the 20% measured on Mistral-translated text, but that figure is a mechanism-specific worst case (an LLM prompted to translate) rather than a representative rate for dedicated NMT, on which the mt-class repair holds (99.3% on held-out mBART-50 documents). We therefore sweep $w \in [0, 1]$, the share of real-world mt content behaving like the worst case. The finding degrades gradually and crosses zero only near $w \approx 0.85$. This bounds an unmeasured quantity; it does not measure it.

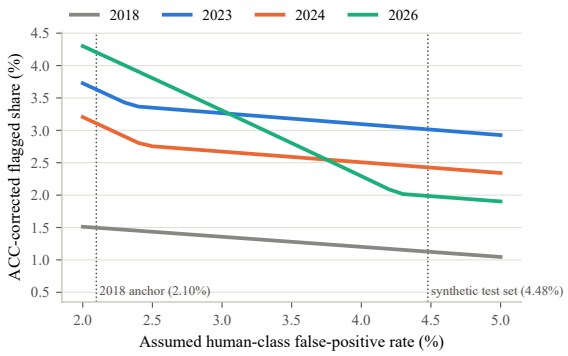


Figure 3: ACC-corrected rate as the assumed human-class FPR sweeps from 2% to 5%. Dotted lines mark the synthetic-test-set (4.48%) and 2018-anchor (2.10%) values. 2018 stays below all three later years throughout; their relative order does not.

A.4 Domain-cluster-robust bootstrap

Each snapshot's domain list is resampled with replacement over 5,000 iterations, routed through the same 2018-anchored ACC correction used for Table 3, with a shared per-iteration confusion-matrix draw across snapshots so paired tests remain valid. Naive-rate intervals widen by 26–55% relative to document-level Wilson intervals; ACC-corrected intervals by 2–5%, because confusion-matrix resampling noise from the 405-document test set already dominates.

A.5 Cross-architecture numbers

BanglaBERT gives 1.08% / 4.40% / 4.01% / 4.90% for 2018 / 2023 / 2024 / 2026 and, like the deployed model, finds 2023-vs-2024 not significant ($p_{\text{holm}} = 0.20$) with 2024-vs-2026 significant ($p_{\text{holm}} = 0.004$). XLM-R-LoRA gives 7.39% / 11.25% / 10.56% / 12.65%, substantially higher in level with the same shape (2023-vs-2024 $p_{\text{holm}} = 0.103$, 2024-vs-2026 $p_{\text{holm}} \approx 0$). Observed three-class κ against the deployed model is 0.20 / 0.26 / 0.28 / 0.28 for BanglaBERT (92.4–97.5% identical

labels) and 0.13 / 0.23 / 0.24 / 0.22 for XLM-R-LoRA (86.2–91.9%), for 2018 / 2023 / 2024 / 2026. Cohen's κ is depressed both by prevalence imbalance and by unequal rater marginals. Fixing each pair's marginals at the rates above, the largest attainable κ is 0.67–0.88 for the deployed model against BanglaBERT, 0.43–0.64 against XLM-R-LoRA and 0.24–0.53 for BanglaBERT against XLM-R-LoRA; we read the observed values against those ceilings rather than against unity.

A.6 Vintage retrains

A frontier model. 24 topic-disjoint training documents authored by Claude Sonnet 5 plus a disjoint 16-document held-out set, produced by direct prompting through the model's chat interface; the sampling configuration was therefore outside our control. *A small local model.* Qwen3 (8B and 4B) proved unusable on a 4GB GPU because its thinking mode consumed the token budget before any output; Llama 3.2 3B instead generated 79 training and 60 held-out documents directly. *Results.* Both retrained detectors classify their held-out documents with 100% accuracy and reproduce the deployed detector's per-snapshot rates to within 0.07 and 0.29 points, with the identical significance pattern (Section 6.4). The ACC correction of Section 6.3 takes its human-class FPR anchor from a vintage-retrained detector: 2.10% of 2018 documents flagged (138 of 6,579), against 2.13% (140) for the deployed detector; the FPR sweep of Figure 3 covers both values.

A.7 Matched-length retrain and rescore

We trained five detectors in which the llm class is replaced entirely by 100 matched-length documents, with human and mt classes drawn as a fresh 100/100 surplus sample per seed, disjoint from the primary benchmark and the calibration holdout. All four snapshots were rescored per seed under a combined leakage exclusion.

Comparison	mt-only sig.	llm-only sig.
2023 vs. 2024	1/5	1/5
2023 vs. 2026	5/5	1/5
2024 vs. 2026	3/5	1/5

Table 6: Seeds (of five) in which each Holm-corrected comparison reaches significance under the matched-length retrain.

The 2018 llm-only rate stays near zero (0.27–0.89%, mean 0.49%). Aggregate mt-only rates are 1.72% $\rightarrow$ 1.64% $\rightarrow$ 2.16% (2023/2024/2026), the original shape; llm-only rates are 0.81% $\rightarrow$ 0.71% $\rightarrow$ 0.81%, essentially flat. A length-binned check is inconclusive at $n \approx 25$–30 per seed, bin and year. Data and scripts: `results/matched_length_retrain_rescore.json`, `scripts/matched_length_retrain_rescore.py`.

A.8 Every method applied to 2018 versus 2026

Table 7 lists every distinct method applied to the 2018-vs-2026 comparison with its primary result.

Check	Method / model	p_{holm}
Raw rate, z-test	Linear (deployed)	≈ 0
Cluster bootstrap	Linear, naive	≈ 0
Cluster bootstrap	Linear, ACC-corr.	0.0032
ACC correction	Linear, 2018 FPR	0.004
Cross-architecture	BanglaBERT	≈ 0
Cross-architecture	XLM-R-LoRA	≈ 0
Per-class rate	mt-only	≈ 0
Per-class rate	llm-only*	6.5×10^{-4}
Vintage retrain	Claude Sonnet 5	≈ 0
Vintage retrain	Llama 3.2 3B	≈ 0
FPR sweep	2%–5%	holds

Table 7: Every distinct method applied to the 2018-vs-2026 comparison, using each method's primary result. These are not independent confirmations: all draw on the same snapshots and several share weights. *Significant in only two of five seeds under the matched-length retrain; every other row holds.

A.9 Targeted `.in` collection

Most of the web uses generic TLDs that do not encode country, so the ccTLD ceiling is largely structural. A Common-Crawl-index-guided fetch targeted WET segments pre-verified to contain `.in` pages (random sampling yields only about 1–4% incidentally). Every year's `.in` cell grew and the rate moved by at most 0.068 points; for 2018, 25 segments each confirmed to contain `.in` content added only two documents. Script: `scripts/tld_covariate_comparison.py`.

A.10 Upstream filtering and domain composition

Re-running the sampling-stage fastText language-ID model on the final corpus gives mean confidence 0.984–0.989 with 2.4–3.4% of documents borderline, stable across snapshots. Between 34.3% and 38.1% of documents have more than 10% non-Bengali characters in their first 2,000 characters, as expected for web text. Restricting the trend to the roughly 90 domains common to all four snapshots

(5.9–8.4% of each) loses significance on most comparisons but keeps the shape; an inverse-propensity-weighted alternative (Kish effective $n = 1{,}978$–$5{,}582$) reaches the same conclusion with substantially more power.

A.11 Regional split: full artifact checks

Cross-tabulating by top-level domain gives an axis native to corpus-building practice. Bangladesh-hosted (`.bd`) content is significantly more flagged than India-hosted (`.in`) content in every testable year: 9.06% against 1.85% in 2023, 10.27% against 3.37% in 2024, 11.11% against 4.19% in 2026, all at $p_{\text{holm}} < 0.001$; 2018 is excluded because its `.in` cell (n=79) cannot power the test. Because the human training class is drawn predominantly from `.bd` sources, three artifact checks preceded any trust in this. **Calibration**: using 2018 as ground truth, false-positive rates are indistinguishable across the two namespaces (1.16% `.bd` against 1.02% `.in`, $p = 0.90$). **Leave-one-domain-out**: dropping each contributing domain in turn preserves the gap across every removal in all three years (533/533, 437/437 and 393/393 iterations significant). **Cluster granularity**: a domain-level test cannot see one operator publishing under several domains, so we re-ran the bootstrap at document, domain and an organisation-proxy granularity that groups domains sharing a registrable base name across TLD suffixes; all nine year-by-granularity combinations remain significant ($p < 0.001$). Support rests on the three together. The scope caveat travels with the finding: documents with a clear country TLD are 10.4% of the pooled corpus, and that subpopulation is shorter and far more government-heavy than the rest (Limitations).

A.12 Onset dating: a reversal

We tried to date the onset by bracketing the ChatGPT and GPT-4 launches with four extra snapshots. The launch-window snapshot, `CC-MAIN-2022-49`, showed a naive spike of 6.24% (3.51% ACC), significantly above its neighbours, and a naive reading would report it as a dateable onset. Class composition ruled out direct ChatGPT authorship at that volume; removing the top three domains moved the rate only to 5.90%, so the spike survived a coarse check. Inverse-propensity domain reweighting (Horvitz–Thompson, weights capped at 10×, Kish effective $n = 2{,}122$–$4{,}856$) dissolved it: 6.24% fell to 4.46% and all pairwise comparisons became non-significant ($p_{\text{holm}} \geq 0.56$). The verdict is a domain-sampling artifact. The transferable point is that narrow-window, single-crawl estimates can manufacture temporal spikes that top-k domain checks miss and inverse-propensity reweighting catches. The four larger trend snapshots show no comparable artifact under identical reweighting.