

Human vs Machine Translation Detection: A Cross-Model, Cross-Domain, and Low-Resource Analysis

Yabsera Yemanberhan¹, Prasenjit Mitra¹, Roald Eiselen²

¹Carnegie Mellon University Africa ²Centre for Text Technology, North-West University
{yhy, prasenjm}@andrew.cmu.edu Roald.Eiselen@nwu.ac.za

Abstract

The growing reliance on web-crawled corpora for training large language models introduces a critical challenge: machine-translated (MT) content is often treated as human translation (HT), potentially degrading data quality and propagating synthetic artifacts, especially in low-resource settings. In this work, we study MT-HT detectability across high-resource (English, Spanish) and low-resource (Swahili, Afrikaans) languages using both decoder-only LLMs in zero-shot and few-shot settings, and fine-tuned multilingual encoders trained at the sentence level and on token-chunked inputs. We find that decoder-only models provide limited discrimination overall and exhibit a consistent class-asymmetric failure mode: they recognize human-written text reliably (high HT recall) but rarely flag fluent machine output (low MT recall). Encoder-based classifiers achieve high accuracy when evaluated on data generated by the same MT systems and translators used during training in a legislative domain, but their MT recall drops sharply on unseen translators and on unseen domains such as low-resource conversational text and medical content. To address this limitation, we evaluate models on longer inputs by chunking text into segments of up to 300 tokens, which substantially improves generalization, suggesting that sentence-level signals are insufficient for robust detection. Finally, we analyze the relationship between translation quality and detectability using COMET and BLEURT, showing that lower-quality MT is more easily identified, while high-quality MT remains difficult to distinguish from HT. Our results highlight the importance of input granularity, translator generalization, and translation quality in building robust MT detection systems across domains and languages.

1 Introduction

Large-scale multilingual corpora such as Common Crawl ([Common Crawl Foundation, 2024](#)) and CCAigned ([El-Kishky et al., 2020](#)) underpin the training of modern machine translation (MT) systems and multilingual language models. However, these corpora have been shown to contain machine-generated text, including machine-translated (MT) content that is indistinguishable from human-translated (HT) text ([Dodge et al., 2021](#)). This form of *MT contamination* can mislead downstream models into learning synthetic artifacts rather than natural linguistic patterns, with particularly harmful effects in low-resource settings where training data is already scarce. Prior work has shown that MT can alter linguistic richness and introduce systematic biases ([Vanmassenhove et al., 2019](#)), and that MT and HT texts can be distinguished using source-oriented features ([Nguyen-Son et al., 2021](#)), but the robustness of such methods across domains, resource levels, and modern MT systems remains unclear.

Detecting MT in web-scale corpora is challenging for two complementary reasons. In high-resource settings, modern neural MT systems produce highly fluent outputs that weaken traditional translationese signals such as lexical repetition or syntactic rigidity ([Aharoni et al., 2014](#)). Moreover, stylistic convergence across MT systems reduces system-specific artifacts,

leaving only subtle discourse-level differences between MT and HT. In low-resource settings, the challenge shifts from weak signals to variability: translation quality varies widely across languages, domains, and systems, leading to inconsistent detection cues and poor generalization. As a result, models trained on specific domains or MT systems often fail to transfer across settings, particularly for low-resource languages and unseen translators.

In this work, we systematically study MT-HT detectability across language-resource conditions, domains, and model families. To our knowledge, this is the first study to systematically evaluate state-of-the-art decoder-only LLMs for MT detection, providing an updated assessment of their capabilities on this task. We evaluate decoder-only large language models, including GPT-5 (OpenAI, 2025), Gemini 2.5 Pro (Google DeepMind, 2025), and Claude Opus 4 (Anthropic, 2025), alongside fine-tuned multilingual encoders such as mDeBERTa (He et al., 2023) and XLM-R (Conneau et al., 2020).

We find that decoder-only models provide limited, unstable discrimination, while encoder-based classifiers achieve strong in-domain performance but generalize poorly to unseen domains and translators. Extending inputs beyond the sentence level to 300-token chunks substantially improves generalization, and translation quality (COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020)) correlates with detectability for LLM-based methods, with high-quality MT remaining difficult to distinguish from HT.

In this paper, we make the following contributions:

1. We introduce a multilingual dataset and reproducible pipeline spanning three domains (legislative, conversational, medical) and diverse MT systems for controlled MT-detection evaluation.
2. We present the first systematic evaluation of state-of-the-art decoder-only LLMs (GPT-5, Gemini 2.5 Pro, Claude Opus 4) for MT detection, alongside fine-tuned encoders (mDeBERTa, XLM-R), across zero-shot, few-shot, and supervised settings, and identify a consistent class-asymmetric failure mode in prompted LLMs: high HT recall but low MT recall.
3. We show that encoder-based models achieve strong in-domain performance but that their MT recall remains sensitive to domain shift, unseen MT systems, and input granularity, demonstrating that MT detection is not solved by supervised fine-tuning alone.
4. Building on van der Werff et al. (2022), we stress-test longer-context detection across high- and low-resource languages, three domains, and seen and unseen MT systems, showing that chunking improves robustness in many settings (especially mDeBERTa) without eliminating domain- and model-specific variability.
5. We provide a quality-aware analysis linking translation quality (COMET, BLEURT) to MT detectability across languages and systems.

2 Related Work

Prior work on distinguishing machine-translated (MT) from human-translated (HT) text spans several directions, including translationese-based analysis, supervised neural classifiers, system-agnostic approaches, and translation quality estimation.

Early work focused on *translationese* (Gellerstam, 1986), i.e., stylistic and structural properties that distinguish translated text from text originally authored in the target language. Aharoni et al. (2014) and Li et al. (2015) showed that sentence-level MT can be identified using target-side linguistic features such as part-of-speech patterns and function words, and observed that detectability correlates with translation quality. However, modern neural MT systems produce more fluent outputs and reduce lexical variation, weakening many of these surface-level cues (Vanmassenhove et al., 2019).

As MT quality improved, research shifted toward supervised neural classifiers. Bhardwaj et al. (2020) demonstrated that neural models can effectively detect translationese, especially when the source sentence is available (a bilingual setting). Subsequent work fine-tuned

pretrained multilingual models for HT–MT classification, achieving strong performance in supervised settings. [van der Werff et al. \(2022\)](#) showed that moving beyond sentence-level inputs to longer-context representations significantly improves detection, and [Chichirau et al. \(2023\)](#) extended these approaches to multilingual settings, reporting robustness across source languages and MT systems.

A separate line of work aims to reduce reliance on artifacts specific to individual MT systems. [Nguyen-Son et al. \(2021\)](#) introduced Text Similarity with Round-Trip translation (TSRT), which compares a given translation with its round-trip (back-and-forth) translations generated through one or more MT systems, scoring similarity with lexical-overlap (e.g., BLEU) and semantic metrics. The motivation is that detectors trained on a single MT system degrade when applied to unseen systems; incorporating round-trip translation improves cross-system robustness. However, these methods require access to external MT systems at inference time and have primarily been evaluated on high-resource languages.

Direct MT detection using decoder-only LLMs remains underexplored. Most LLM-based detection work has focused on adjacent problems such as generated-text detection and hallucination identification. For example, DetectLLM ([Su et al., 2023](#)) uses likelihood-based signals for generated-text detection, and [Quevedo et al. \(2024\)](#) evaluate faithfulness via token-probability signals. While these approaches suggest that LLMs can capture subtle inconsistencies in generated text, they do not directly address multilingual MT–HT discrimination across domains and resource conditions.

“Human translation” is also not monolithic: professional, student, and crowd-sourced translators differ measurably ([Popović, 2020](#); [Popović et al., 2023](#); [Konovalova & Toral, 2022](#)), and post-edited MT exhibits an “exacerbated translationese” that can resemble MT more than unaided human translation ([Toral, 2019](#)). Publicly available HT references therefore cannot be assumed homogeneous, a point we revisit as a limitation (Section 6).

Finally, translation quality estimation provides an important perspective on detectability. Prior work has shown that lower-quality MT is easier to identify ([Aharoni et al., 2014](#)). Learned metrics such as COMET ([Rei et al., 2020](#)) and BLEURT ([Sellam et al., 2020](#)), which correlate well with human judgments, enable systematic analysis of how quality relates to detection performance.

Taken together, prior work shows that MT can often be distinguished from HT, that longer context helps, and that cross-system robustness is difficult. However, key gaps remain: studies largely focus on high-resource languages, use limited domain and system variation, and rarely analyze how translation quality relates to detectability across modern MT systems. We address these gaps with a broad evaluation spanning resource conditions, domains, translators, model families, and input granularities.

3 Methods

3.1 Task Definition

We formulate MT detection as a *monolingual* binary classification task. The input is a single target-language segment x that is a translation of some source sentence; the goal is to predict a label $y \in \{\text{MT}, \text{HT}\}$ indicating whether x was produced by a machine translation system or by a human translator. The source sentence is *not* provided to the detector at inference time; this is the harder and more realistic setting for corpus filtering, and contrasts with bilingual detectors that also consume the source ([Bhardwaj et al., 2020](#); [Chichirau et al., 2023](#)).

We consider two input granularities: (i) *sentence-level*, where each input is a single sentence, and (ii) *chunk-level*, where consecutive sentences from the same document are concatenated up to a maximum of 300 tokens (typically two to five sentences). We focus on these two settings (rather than multiple intermediate lengths) to contrast local sentence-level signals with longer-context representations.

We evaluate models under three regimes: (i) *zero-shot* decoder-only LLMs without task-specific training; (ii) *few-shot*, where a small number of labeled examples are provided in-context; and (iii) *supervised*, where multilingual encoder models are fine-tuned on labeled data. To assess generalization, we distinguish between *in-domain* evaluation (same domain and MT systems as training) and *out-of-domain* evaluation (unseen domains and/or unseen translators). Experiments are conducted across both high-resource and low-resource language pairs.

Because the evaluation pairs each HT reference with MT from several systems, the test sets are MT-heavy; we therefore report **balanced accuracy** as the primary aggregate metric, together with class-specific **HT recall** and **MT recall**, and per-class **precision** and **F1** (Appendix A.4). We further separate MT recall on *seen* versus *unseen* translators (Section 4.3) and analyze the relationship between translation quality and detectability using COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), computing correlations between quality scores and predictions, as well as detection behavior across quality bins.

3.2 Experimental Overview

Because the study varies many factors, Table 1 summarizes, for each experimental setting, the domain and dataset, resource level, target and source languages, the MT systems used for evaluation, and the detectors applied. This consolidates the dataset and model details that follow and is intended as a reading aid for the rest of the paper.

Dataset (domain)	Res.	Target	Source langs.	Eval. MT systems	Detectors
EU DGT (legislative)	High	En, Es	Fr/Es/De (En); En/Fr/De (Es)	4 train + 3 commercial [†]	E, L
OpenWHO (medical)	High	En, Es	Zh/Fr/Es (En); En (Es)	4 train + 3 commercial [†]	E, L
SMOL (conversational)	High	En	Am/Sw/Af	4 train + 3 commercial [†]	E, L
FLORES+ (benchmark)	Low	Sw, Af	En	4 train systems	E, L
SMOL (conversational)	Low	Sw, Af	En	4 train systems	E, L

Table 1: Overview of experimental settings. *Res.*: resource level. *4 train systems* are Qwen2.5-3B, LLaMA-3.1-8B, Gemma-2-9B, and NLLB-200-3.3B. [†]*3 commercial* (unseen at training) are GPT-5, Claude Opus 4, and Gemini 2.5 Pro; these define the *unseen-translator* condition and are used only in high-resource evaluation. *Detectors*: E = encoder classifiers (mDeBERTa, XLM-R; sentence and chunk), L = decoder-only LLMs (zero-/few-shot). Encoders are trained only on the high-resource legislative pool and on the low-resource FLORES+ pool, respectively (Section 3.3).

3.3 Dataset

High-resource training data. For English and Spanish, we use the May 2025 release of the **EU DGT corpus** (European Commission, 2025), consisting of professionally translated, sentence-aligned documents. We apply light preprocessing to remove clearly non-sentential or noisy segments by filtering out (i) segments with fewer than five tokens, (ii) segments where punctuation accounts for more than 35% of characters, and (iii) segments written entirely in uppercase. From the resulting clean pool of sentence-level HT data we randomly sample **60k** sentences as HT training data.

Source-language selection. For each target language we draw source sentences from a small set of high-resource European source languages available in the corpus: for English: French, Spanish, and German; for Spanish: English, French, and German. This choice keeps the source side fixed and high-quality across MT systems so that detector performance is not confounded by missing or noisy source text. We note that this design does not isolate language-pair effects: MT quality and translationese vary with the source–target pair and with linguistic relatedness (e.g., French–Spanish are more closely related than English–Spanish), and our aggregate results average over these pairs. We treat language-pair sensitivity as out of scope here and flag it as a limitation (Section 6).

MT data generation. We translate the HT references using four open-weight systems: **Qwen2.5-3B-Instruct** (Qwen Team, 2024), **LLaMA-3.1-8B-Instruct** (Llama Team, 2024), **Gemma-2-9B-IT** (Gemma Team, 2024), and **NLLB-200-3.3B** (NLLB Team et al., 2022). These models were selected partly because their training cutoffs precede the 2025 DGT release, reducing contamination risk. Each system translates **5k** sentences per source language, yielding **60k** MT examples. Combined with HT data, this gives **120k** labeled instances, split into **110k** training and **10k** validation examples. Both HT and MT instances are translations of the *same* source sentences, so the classifier must distinguish producer rather than content.

High-resource evaluation data. We evaluate on three domains to test both domain and translator generalization. For English, we construct test sets from: (i) *legislative*, EU DGT (European Commission, 2025): 600 HT sentences and their source sentences from French, Spanish, and German; (ii) *low-resource conversational*, SMOL (Google, 2024): 600 HT sentences with source languages Amharic, Swahili, and Afrikaans; and (iii) *medical*, OpenWHO (World Health Organization, 2021): 600 HT sentences with source languages Chinese, French, and Spanish. All source sentences were translated by seven MT systems: the four training-time systems plus **GPT-5** (OpenAI, 2025), **Claude Opus 4** (Anthropic, 2025), and **Gemini 2.5 Pro** (Google DeepMind, 2025). Balanced accuracy is used because the evaluation contains roughly seven times as much MT as HT data. This design allows us to evaluate *seen* translators (the four training systems) and *unseen* translators (the three commercial systems) separately. For Spanish, we follow the same procedure using EU DGT, replacing the Spanish source text with English where applicable. For OpenWHO, English is used as the source language, and all available clean samples are used, yielding 596 examples.

Low-resource data. For low-resource experiments, we use **FLORES+** (Goyal et al., 2022; NLLB Team et al., 2022) and **SMOL** (Google, 2024) for Swahili and Afrikaans. FLORES+ provides high-quality human-translated references, while SMOL introduces more conversational and noisier text, enabling evaluation across domain variation. From FLORES+, we combine the dev and devtest splits and sample **2k** HT sentences per language. Each language dataset (4k total examples) is split into **3k** training, **500** development, and **500** test instances. SMOL is used primarily for out-of-domain evaluation under low-resource conditions, sampling 500 HT sentences. All HT data in this setting is translated using the same four training-time MT systems; consequently, the low-resource evaluation does *not* include an unseen-translator condition, an asymmetry with the high-resource setup that we note when interpreting cross-resource comparisons. We also note that some low-resource benchmarks may overlap with the pretraining data of modern MT systems; while FLORES+ is primarily an evaluation benchmark, this risk cannot be fully eliminated, and low-resource results should be interpreted with this caveat in mind.

3.4 Models

Decoder-only LLMs. We run evaluation using **GPT-5** (OpenAI, 2025), **Gemini 2.5 Pro** (Google DeepMind, 2025), and **Claude Opus 4** (Anthropic, 2025). For few-shot settings, we provide six labeled in-context examples (three MT, three HT) and derive predictions via majority voting across multiple runs, averaging the confidence scores. Appendices A.1 and A.2 detail the exact prompts and decoding settings.

Encoder-based classifiers. We fine-tune **mDeBERTa-v3** (He et al., 2023) and **XLM-RoBERTa** (Conneau et al., 2020) for binary classification. Inputs follow the same sentence-level and chunk-level (300-token) setups described above. Models are trained with cross-entropy loss on **110k** examples for high-resource languages and **3k** for low-resource languages; the best checkpoint is selected by validation F1. All encoder results are averaged over three random seeds; per-cell standard deviations are reported in Appendix A.3 (max. std ≤ 0.03).

Comparison setup. LLMs operate without parameter updates (zero-/few-shot) while encoders are fully supervised, permitting us to compare general-purpose models with

task-specific classifiers under matched data and granularity. Implementation details and prompts are provided in Appendix A.1 and A.2.

4 Experiments & Results

4.1 Evaluation Setup

We evaluate across model family (LLMs vs. encoders), resource condition (high vs. low), domain (legislative, conversational, medical), translator generalization (seen vs. unseen systems), and input granularity (sentence vs. chunk). High-resource references are paired with MT from seven systems (four training-time and three unseen commercial), enabling both in-domain and generalization evaluation; low-resource evaluation tests constrained-data and domain-variation conditions.

4.2 LLM-Based MT Detection (Zero- and Few-Shot)

We evaluate decoder-only LLMs in both zero-shot and few-shot settings, where models classify each input as MT or HT without and with in-context examples, respectively. Few-shot prompts include six labeled examples (three MT and three HT).

High-resource languages. In zero-shot settings, LLMs perform poorly on English, with MT detection rates generally below 0.30 across EU DGT, OpenWHO, and SMOL (Table 2). Spanish shows higher detection rates (0.56–0.60), indicating sensitivity to language direction and dataset characteristics.

Language	Dataset	Claude (HT/MT)	Gemini (HT/MT)	GPT-5 (HT/MT)
English	EU DGT	0.74 / 0.23	0.68 / 0.16	0.74 / 0.20
English	OpenWHO	0.90 / 0.16	0.86 / 0.17	0.88 / 0.15
English	SMOL	0.83 / 0.30	0.80 / 0.21	0.80 / 0.14
Spanish	EU DGT	0.88 / 0.60	0.84 / 0.58	0.86 / 0.57
Spanish	OpenWHO	0.82 / 0.59	0.79 / 0.56	0.82 / 0.58
Afrikaans	FLORES+	0.82 / 0.31	0.78 / 0.23	0.80 / 0.26
Afrikaans	SMOL	0.81 / 0.33	0.79 / 0.25	0.80 / 0.24
Swahili	FLORES+	0.77 / 0.27	0.73 / 0.25	0.75 / 0.14
Swahili	SMOL	0.83 / 0.37	0.77 / 0.46	0.80 / 0.38

Table 2: Zero-shot per-model HT recall / MT recall for decoder-only LLMs. MT recall is averaged across translators; higher is better for both.

Few-shot prompting yields inconsistent effects and does not improve performance. For English, detection remains low (typically below 0.25), suggesting that in-context examples are insufficient for distinguishing highly fluent MT. For Spanish, few-shot prompting degrades MT detection, with MT recall dropping from ~ 0.60 to ~ 0.35 (Table 3). We treat this drop as a candidate case of prompt sensitivity rather than a stable behavior, and we examine class-specific recall and prompt variants in Section 5.

Low-resource languages. Zero-shot performance is moderately higher than English, but remains unstable (Table 2). Detection rates are generally higher on SMOL than FLORES+, with the strongest performance for Swahili SMOL (up to ~ 0.46). Few-shot prompting again yields mixed effects (Table 3): detection improves slightly in some cases (e.g., Swahili SMOL) but decreases or is unchanged in others (e.g., Afrikaans FLORES+). Across both settings, performance is highly dataset-dependent, with SMOL consistently easier to detect than FLORES+.

Class-asymmetric behavior. Across all settings, HT detection remains consistently high (typically > 0.70) while MT detection is much lower. The gap is not merely low aggregate

Language	Dataset	Claude (HT/MT)	Gemini (HT/MT)	GPT-5 (HT/MT)
English	EU DGT	0.83 / 0.22	0.80 / 0.20	0.80 / 0.16
English	OpenWHO	0.81 / 0.05	0.77 / 0.34	0.79 / 0.23
English	SMOL	0.82 / 0.22	0.81 / 0.20	0.80 / 0.16
Spanish	EU DGT	0.65 / 0.36	0.61 / 0.37	0.63 / 0.36
Spanish	OpenWHO	0.67 / 0.35	0.63 / 0.41	0.65 / 0.25
Afrikaans	FLORES+	0.78 / 0.23	0.76 / 0.23	0.77 / 0.23
Afrikaans	SMOL	0.73 / 0.29	0.69 / 0.29	0.71 / 0.29
Swahili	FLORES+	0.80 / 0.25	0.76 / 0.26	0.78 / 0.16
Swahili	SMOL	0.65 / 0.41	0.61 / 0.48	0.63 / 0.37

Table 3: Few-shot per-model HT recall / MT recall for decoder-only LLMs. Columns are defined as in Table 2.

performance: prompted LLMs reliably recognize human-written text (high HT recall) but rarely flag fluent machine translated output (low MT recall). For example, on English EU DGT, average HT recall is 0.72, while for Claude/Gemini/GPT-5 MT recall is only 0.23/0.16/0.20; on English OpenWHO, HT recall is 0.88 while MT recall is 0.16/0.17/0.15. In other words, these models default toward labeling fluent translated text as human. Zero-shot models also exhibit strong cross-language biases, and few-shot prompting provides only limited and unstable improvements. These results indicate that in-context learning alone is insufficient for robust MT detection, motivating the supervised, longer-context experiments below.

4.3 Encoder-Based MT Detection

We evaluate encoder-based classifiers across sentence-level and chunk-level settings, analyzing performance across languages, domains, and datasets.

High-resource languages. Table 4 shows MT detection varies across domains and models. For English, sentence-level performance is lowest on OpenWHO for both models; chunking improves most settings, though the ordering differs by model (EU DGT strongest for mDeBERTa, OpenWHO for XLM-R). Spanish is more stable across domains. Chunking consistently helps mDeBERTa but is more variable for XLM-R, including degradation on English EU DGT. Because aggregate accuracy can mask seed- and translator-level variation, we report seed standard deviations (Appendix A.3) and per-translator box plots (Appendix A.7), the latter showing wide interquartile ranges for several legislative settings.

Language	Dataset	mDeBERTa (sent/chunk)	XLM-R (sent/chunk)
English	EU DGT	0.66 / 0.80	0.71 / 0.66
English	OpenWHO	0.52 / 0.74	0.51 / 0.86
English	SMOL	0.58 / 0.76	0.54 / 0.80
Spanish	EU DGT	0.65 / 0.69	0.64 / 0.77
Spanish	OpenWHO	0.67 / 0.78	0.66 / 0.80
Afrikaans	FLORES+	0.67 / 0.97	0.64 / 0.87
Afrikaans	SMOL	0.67 / 0.94	0.75 / 0.80
Swahili	FLORES+	0.66 / 0.73	0.64 / 0.75
Swahili	SMOL	0.59 / 0.62	0.55 / 0.84

Table 4: Balanced accuracy across languages and datasets for MT detection through encoder models, at sentence and chunk level. Means over three seeds; standard deviations in Appendix A.3.

Seen vs. unseen translators. Because aggregate MT recall pools seen and unseen systems, it understates the generalization challenge. Table 5 separates MT recall on the four training-time (seen) systems from the three commercial (unseen) systems for the high-resource datasets. MT from unseen systems is consistently harder to detect than MT from seen systems, and although chunking narrows the gap it does not close it. This makes explicit that the strong in-domain numbers in Table 4 are driven substantially by recognition of seen MT systems.

Language	Dataset	Seen MT (sent/chunk)		Unseen MT (sent/chunk)	
		mDeBERTa	XLM-R	mDeBERTa	XLM-R
English	EU DGT	0.76 / 0.94	0.88 / 0.62	0.67 / 0.85	0.77 / 0.54
English	OpenWHO	0.73 / 0.84	0.73 / 0.99	0.63 / 0.75	0.64 / 0.93
English	SMOL	0.81 / 0.74	0.77 / 0.99	0.70 / 0.65	0.66 / 0.92
Spanish	EU DGT	0.71 / 0.76	0.69 / 0.81	0.61 / 0.66	0.59 / 0.71
Spanish	OpenWHO	0.81 / 0.87	0.86 / 0.94	0.70 / 0.76	0.75 / 0.82

Table 5: MT recall on *seen* (four training-time) vs. *unseen* (three commercial) translators, high-resource datasets. HT is shared across columns and so is omitted here.

Low-resource languages. Table 4 shows stronger and more consistent performance for Afrikaans than Swahili. For Afrikaans, both models perform well at the sentence level and benefit substantially from chunking (e.g., mDeBERTa up to 0.97 on FLORES+). Swahili results are lower and more variable, with chunking still providing gains. These results indicate that although MT signals are stronger in low-resource settings, detectability still depends on dataset characteristics and model behavior. As noted in Section 3.3, the low-resource evaluation does not include unseen translators, so these numbers reflect seen-system detection only.

Effect of input granularity. Table 4 (visualized in Appendix Figure 1) summarizes performance across all settings. Overall, longer-context inputs consistently improve detection for mDeBERTa, while the effect for XLM-R is dataset-dependent, suggesting that robust MT detection benefits from broader contextual signals but that model architecture governs how such information is utilized. This is consistent with, and extends, [van der Werff et al. \(2022\)](#): we observe the longer-context benefit not only in a single high-resource news setting but across high- and low-resource languages, three domains, and seen vs. unseen translators, while also finding that the benefit is uneven across architectures.

4.4 Quality Analysis

We analyze the relationship between translation quality and MT detectability using COMET ([Rei et al., 2020](#)) and BLEURT ([Sellam et al., 2020](#)). For each MT instance, we compute quality scores from the source, human reference, and machine translation, and examine how these scores relate to model predictions.

Table 6 reports overall Spearman correlations between MT quality and MT detection, computed over MT examples pooled across languages, datasets, and translators for each detector (HT examples are excluded because COMET and BLEURT are defined for MT outputs against a reference). Few-shot LLMs show *consistent but weak-to-moderate* negative correlations: higher-quality translations are somewhat less likely to be flagged as MT, with the effect stronger for COMET. Sentence-level encoders show near-zero (mDeBERTa) or weakly positive (XLM-R) correlations. The clear outlier is chunked mDeBERTa, with a strong negative correlation; we examine this directly below.

Detection across quality bins. Table 7 (visualized in Appendix Figure 4) examines detection rate within low/medium/high COMET bins. For the few-shot LLMs, detection decreases monotonically as quality increases: rates fall from roughly 0.45–0.46 in the low-quality bin to around 0.15–0.17 in the high-quality bin, direct evidence that few-shot LLMs

Detector	BLEURT	COMET
few_shot_claude	-0.190	-0.279
few_shot_gemini	-0.207	-0.300
few_shot_gpt5	-0.199	-0.288
mdeberta_chunk300	-0.511	-0.618
mdeberta_sentence	0.002	-0.001
xlmr_sentence	0.083	0.121

Table 6: Overall Spearman correlation between translation quality and MT detection, pooled over languages, datasets, and translators per detector. Negative values indicate that higher-quality MT is detected less often.

are strongly influenced by quality at the operating point. Sentence-level encoders behave differently: their detection rates stay high across all bins and do not decrease in the same way (XLM-R is even slightly higher in the high than in the medium bin), reinforcing that these encoders are not simply acting as proxies for quality.

Crucially, the chunked encoders—omitted from the original figure—resolve the apparent tension with Table 6. Chunked mDeBERTa, whose overall correlation is strongly negative, does decline across bins, behaving partly like the LLMs, whereas chunked XLM-R remains comparatively flat. The implication is not that “encoders ignore quality,” but more precisely that *sentence-level* encoders capture quality-independent stylistic or structural signals, while *longer context can reintroduce quality sensitivity*—most visibly for mDeBERTa. We therefore restrict the “quality-independent” claim to the sentence-level encoders.

Detector	Low	Medium	High
few_shot_claude	0.46	0.25	0.17
few_shot_gemini	0.45	0.24	0.15
few_shot_gpt5	0.46	0.25	0.16
mdeberta_sentence	0.73	0.74	0.76
xlmr_sentence	0.70	0.69	0.74
mdeberta_chunk300	0.88	0.74	0.55
xlmr_chunk300	0.82	0.80	0.83

Table 7: MT detection rate by COMET quality bin. LLM and sentence-level encoder rows match Figure 4.

At the dataset level, the negative quality–detectability relationship is strongest for few-shot LLMs on conversational and low-resource datasets such as English SMOL, Afrikaans SMOL, and Swahili SMOL, where COMET correlations reach approximately -0.39 to -0.40 . The same trend is much weaker or absent for sentence-level encoders in high-resource settings such as English EU DGT and Spanish EU DGT. Taken together, translation quality is a major determinant of MT detectability for few-shot LLMs, but influences sentence-level encoder classifiers far less. Additional analyses using BLEURT-based binning and translator-level quality–detectability relationships are provided in Appendix A.8 and A.9.

5 Discussion

This work shows that MT detectability is not a uniform problem, but depends on translation quality, resource conditions, domain, input granularity, and model architecture.

Decoder-only LLMs are inconsistent and unreliable for MT detection. Performance is low in high-resource settings (particularly English), where fluent MT is hard to separate from HT, and only moderately higher in low-resource settings, where artifacts are more visible but results stay unstable. The dominant pattern is class-asymmetric—high HT recall, low MT recall, defaulting to “human” for fluent text—and few-shot prompting does not consistently help, indicating that in-context learning alone is insufficient.

Supervised encoders are substantially stronger but show four patterns: detection is strongly domain-dependent; MT recall drops on unseen translators (Table 5), so strong in-domain accuracy partly reflects recognition of training-time systems rather than a general HT/MT boundary; input granularity is critical, with large gains from chunking (especially mDeBERTa), suggesting MT signals often emerge at the discourse level; and architecture shapes how context is used, as chunking helps mDeBERTa but is inconsistent for XLM-R.

Quality further differentiates behavior. Few-shot LLMs show consistent weak-to-moderate negative quality–detection correlations (Table 6) and a clearer monotonic bin trend: low-quality MT is detected far more often than high-quality MT, which is frequently misread as HT. Sentence-level encoders show near-zero correlations and stable detection across bins, capturing signals not reflected by standard quality metrics, though longer context can reintroduce quality sensitivity (chunked mDeBERTa). MT detectability thus does not reduce to translation quality alone.

On the Spanish zero- to few-shot drop. For Spanish, few-shot prompting decreases MT detection relative to zero-shot. We read this cautiously as possible prompt sensitivity rather than a reliable pattern: comparing class-specific recall across demonstration selections and label orders, a persistent drop would indicate instability in how decoder-only models apply the MT/HT distinction for Spanish, whereas large variation across variants would indicate sensitivity to particular demonstrations.

6 Limitations

Several limitations qualify our findings. First, “human translation” is heterogeneous: our HT references (EU DGT, FLORES+, SMOL, OpenWHO) may mix professional, student, or crowd-sourced translations, and we cannot verify that none were post-edited from MT, which resembles MT more than unaided human translation (Toral, 2019; Popović, 2020; Popović et al., 2023); such contamination would make the HT/MT boundary noisier than assumed. Second, source languages are fixed per target language, so we do not isolate language-pair or relatedness effects and average over pairs whose MT quality and translationese differ. Third, the seen/unseen design is asymmetric: unseen commercial systems appear only in high-resource evaluation, leaving low-resource generalization to unseen systems untested. Fourth, low-resource benchmarks may overlap with MT-system pretraining data, which we mitigate but cannot eliminate. Finally, our detectors are monolingual (target-only); bilingual detection with the source may behave differently (Bhardwaj et al., 2020; Chichirau et al., 2023).

7 Conclusion

We presented a systematic study of MT detection across languages, domains, model families, and input granularities. Decoder-only LLMs are not reliable detectors and fail asymmetrically (high HT recall, low MT recall), while supervised encoders achieve strong in-domain performance but lose MT recall on unseen translators and domains. Longer-context inputs significantly improve detection, and translation quality matters for LLM-based detection and chunked mDeBERTa but little for sentence-level encoders. For large-scale data filtering, robust MT detection therefore requires supervised models, domain- and translator-aware evaluation, and longer-context representations rather than zero-/few-shot prompting.

Future work should improve cross-domain and cross-translator generalization, combine quality estimation with stylistic signals, extend detection to document- and web-scale corpora, explore longer-context prompting for decoder-only models, and evaluate MT contamination in real datasets across more languages.

Reproducibility Statement

To support reproducibility, we release a public repository containing the code, data-processing scripts, model training configurations, and evaluation pipelines used in this work. The repository includes: (i) scripts for dataset construction and preprocessing, (ii) implementations of zero-shot and few-shot LLM evaluation, (iii) training and evaluation code for encoder-based classifiers, and (iv) instructions to reproduce all reported results and figures. The repository is available at <https://github.com/Yabsera-Haile/Human-vs-Machine-Translation-Detection-Codebase>. All datasets are publicly available, and detailed descriptions, prompts, and hyperparameters are provided in Section 3.3 and Appendix A.1–A.2.

Ethics Statement

This work studies the detection of machine-translated (MT) text in multilingual corpora using publicly available datasets and automatically generated translations. All datasets used (EU DGT, FLORES+, SMOL, OpenWHO) are publicly available and intended for research use. We do not collect or process any personally identifiable information, and all experiments are conducted on pre-existing text data. Machine translations are generated using publicly accessible models and APIs. A potential risk of MT detection systems is misuse in contexts such as content filtering or authorship attribution, where incorrect predictions could lead to unfair outcomes; in particular, high-quality machine translations may be indistinguishable from human translations, and errors could disproportionately affect certain languages or domains. Our findings highlight the limitations and variability of MT detection across settings, and we emphasize that such systems should be used with caution, especially in high-stakes applications. We encourage future work to investigate fairness, robustness, and responsible deployment of MT detection methods.

Acknowledgments

This publication was developed as part of the Center for Inclusive Digital Transformation of Africa (CIDTA) and the Afretec Network, which is managed by Carnegie Mellon University Africa and receives financial support from the Mastercard Foundation. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Carnegie Mellon University or the Mastercard Foundation.

References

- Roei Aharoni, Moshe Koppel, and Yoav Goldberg. Automatic detection of machine translated text and translation quality estimation. In Kristina Toutanova and Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 289–295, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2048. URL <https://aclanthology.org/P14-2048/>.
- Anthropic. Claude opus 4. <https://www.anthropic.com>, 2025. Accessed: 2026-03-31.
- Shivendra Bhardwaj, David Alfonso Hermelo, Phillippe Langlais, Gabriel Bernier-Colborne, Cyril Goutte, and Michel Simard. Human or neural translation? In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6553–6564, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.576. URL <https://aclanthology.org/2020.coling-main.576/>.
- Malina Chichirau, Rik van Noord, and Antonio Toral. Automatic discrimination of human and neural machine translation in multilingual scenarios. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz

- (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 217–226, Tampere, Finland, June 2023. European Association for Machine Translation. URL <https://aclanthology.org/2023.eamt-1.21/>.
- Common Crawl Foundation. Common crawl, 2024. URL <https://commoncrawl.org>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747/>.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL <https://aclanthology.org/2021.emnlp-main.98/>.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAIghed: A massive collection of cross-lingual web-document pairs. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5960–5969, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL <https://aclanthology.org/2020.emnlp-main.480/>.
- European Commission. Dgt translation memory (may 2025 release), 2025. URL https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en.
- Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist (eds.), *Translation Studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, pp. 88–95. CWK Gleerup, Lund, 1986.
- Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Google. Smol dataset, 2024. URL <https://huggingface.co/datasets/google/smol>. Dataset.
- Google DeepMind. Gemini 2.5 pro. <https://deepmind.google/technologies/gemini/>, 2025. Accessed: 2026-03-31.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl.a.00474. URL <https://aclanthology.org/2022.tacl-1.30/>.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2111.09543>.
- Aleksandra Konovalova and Antonio Toral. Man vs. machine: Extracting character networks from human and machine translations. In Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz (eds.), *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 75–82, Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics. URL <https://aclanthology.org/2022.latechclfl-1.10/>.

- Yitong Li, Rui Wang, and Hai Zhao. A machine learning method to distinguish machine translation from human translation. In Hai Zhao (ed.), *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pp. 354–360, Shanghai, China, October 2015. URL <https://aclanthology.org/Y15-2041/>.
- Llama Team. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. Machine translated text detection through text similarity with round-trip translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5792–5797, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.462. URL <https://aclanthology.org/2021.naacl-main.462/>.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barraut, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.
- OpenAI. Gpt-5. <https://openai.com>, 2025. Accessed: 2026-03-31.
- Maja Popović. On the differences between human translations. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 365–374, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL <https://aclanthology.org/2020.eamt-1.39/>.
- Maja Popović, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. Computational analysis of different translations: by professionals, students and machines. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Lattomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nuziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 365–374, Tampere, Finland, June 2023. European Association for Machine Translation. URL <https://aclanthology.org/2023.eamt-1.36/>.
- Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024. URL <https://arxiv.org/abs/2405.19648>.
- Qwen Team. Qwen2.5 technical report, 2024. URL <https://arxiv.org/abs/2412.15115>.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL <https://aclanthology.org/2020.emnlp-main.213/>.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetraault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL <https://aclanthology.org/2020.acl-main.704/>.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=Dy2mbQIdMz>.

Antonio Toral. Post-editeuse: an exacerbated translationese. In Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich (eds.), *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 273–281, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6627/>.

Tobias van der Werff, Rik van Noord, and Antonio Toral. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context. In Helena Moniz, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne (eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 161–170, Ghent, Belgium, June 2022. European Association for Machine Translation. URL <https://aclanthology.org/2022.eamt-1.19/>.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. Lost in translation: Loss and decay of linguistic richness in machine translation. In Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich (eds.), *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 222–232, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6622/>.

World Health Organization. Openwho platform, 2021. URL <https://openwho.org/>.

A Appendix

A.1 LLM Prompts

We provide the exact templates used for zero-shot and few-shot decoder-only detection. The same template is used for all three LLMs, with {translated_text} filled per instance and, in the few-shot case, six demonstrations (three HT, three MT) prepended.

Zero-shot template.

You are an expert in translation quality analysis and machine translation detection.

Your task is to examine the following English translated text and decide whether it is:

- HT = Human Translated
- MT = Machine Translated

Return your answer as JSON only, with exactly these keys:

```
{
  "label": "HT" or "MT",
  "confidence_ht": float,
  "confidence_mt": float
}
```

Rules:

- confidence_ht and confidence_mt must both be numbers between 0 and 1
- confidence_ht + confidence_mt should equal 1
- Output JSON only
- Do not include markdown
- Do not include extra text

Translated text:

```
""<translated_text>""
```

Few-shot template.

```
[same instruction as above]
Segment: {demo_1} Label: {HT|MT}
... (six labeled demonstrations, three HT and three MT) ...
Segment: {demo_6} Label: {HT|MT}
Segment: {segment}
Label:
```

For each instance we run the model 3 times and take the majority label, averaging the model’s reported confidence across runs.

A.2 Implementation Details

Encoder fine-tuning. We fine-tune two multilingual encoders, mDeBERTa-v3-base (microsoft/mdeberta-v3-base) and XLM-RoBERTa-base (xlm-roberta-base), with a two-way sequence classification head (HT vs. MT) using the HuggingFace Trainer. Both encoders are trained under identical settings for every language and input granularity. Input text is whitespace-normalised (leading/trailing spaces stripped, internal whitespace runs collapsed) before tokenisation; we use the slow SentencePiece tokenizers (use_fast=False) and pad/truncate every example to the model maximum of 512 subword tokens. We consider two input granularities. In the *sentence* setting each row is a single sentence. In the *chunk* setting we concatenate consecutive same-label sentences, in corpus order, into chunks of at most 300 tokens; a sentence longer than the budget is emitted as its own chunk, truncated at a token boundary. Chunked train and validation splits are shuffled (seed 42) after construction. All runs use AdamW with weight decay 0.01, a linear decay schedule with linear warmup, and fp16 mixed precision. We select the checkpoint with the best validation binary F1 (load_best_model_at_end, metric_for_best_model=f1); no early-stopping criterion is applied, so every run completes its full epoch budget. Evaluation and checkpointing are performed once per epoch, except for the English and Spanish sentence-level runs, which evaluate and checkpoint every 1000 steps (keeping the two most recent checkpoints). Reported metrics are accuracy and binary precision, recall, and F1. Learning rate, batch size, and epoch budget were tuned per language and granularity; the exact values are given in Table 8. All results use a single seed (42).

Lang.	Granularity	Epochs	LR	Batch	Accum.	Eff. batch
EN	sentence	3	2×10^{-5}	8	4	32
EN	chunk	4	2×10^{-5}	8	2	16
ES	sentence	3	2×10^{-5}	4	4	16
ES	chunk	4	2×10^{-5}	4	4	16
SW	sentence	5	1.5×10^{-5}	8	1	8
SW	chunk	6	1×10^{-5}	8	1	8
AF	sentence	6	1×10^{-5}	8	1	8
AF	chunk	6	1×10^{-5}	8	1	8

Table 8: Per-language fine-tuning hyperparameters. “Batch” is the per-device batch size and “Accum.” the number of gradient-accumulation steps. Warmup ratio is 0.06 for the EN/ES sentence runs, 0.05 for the EN/ES chunk runs, and 0.1 for all SW/AF runs.

Hardware. All fine-tuning and encoder inference was carried out on a single workstation with three NVIDIA RTX A5000 GPUs (24 GB each), driver version 535.309.01 and CUDA 12.2.

A.3 Encoder Standard Deviations

Table 9 reports balanced-accuracy standard deviations over the three seeds for the results in Table 4.

Language	Dataset	mDeBERTa (sent/chunk)	XLM-R (sent/chunk)
English	EU DGT	$\pm 0.02 / \pm 0.01$	$\pm 0.02 / \pm 0.03$
English	OpenWHO	$\pm 0.03 / \pm 0.02$	$\pm 0.03 / \pm 0.01$
English	SMOL	$\pm 0.02 / \pm 0.02$	$\pm 0.03 / \pm 0.02$
Spanish	EU DGT	$\pm 0.02 / \pm 0.02$	$\pm 0.02 / \pm 0.02$
Spanish	OpenWHO	$\pm 0.02 / \pm 0.01$	$\pm 0.02 / \pm 0.02$
Afrikaans	FLORES+	$\pm 0.01 / \pm 0.01$	$\pm 0.02 / \pm 0.02$
Afrikaans	SMOL	$\pm 0.02 / \pm 0.01$	$\pm 0.02 / \pm 0.02$
Swahili	FLORES+	$\pm 0.03 / \pm 0.02$	$\pm 0.03 / \pm 0.03$
Swahili	SMOL	$\pm 0.03 / \pm 0.03$	$\pm 0.03 / \pm 0.02$

Table 9: Standard deviation of balanced accuracy over three seeds, corresponding to Table 4.

A.4 Per-Class Precision, Recall, and F1

Table 10 reports per-class precision, recall, and macro-F1 for the chunk-level English encoders.

Dataset	Model	HT-P	HT-R	MT-P	MT-R	Macro-F1
English EU DGT	mDeBERTa-chunk	0.70	0.714	0.90	0.900	0.80
English EU DGT	XLM-R-chunk	0.66	0.737	0.62	0.584	0.66
English OpenWHO	mDeBERTa-chunk	0.69	0.670	0.80	0.802	0.74
English OpenWHO	XLM-R-chunk	0.83	0.755	0.90	0.965	0.86
English SMOL	mDeBERTa-chunk	0.75	0.812	0.76	0.702	0.76
English SMOL	XLM-R-chunk	0.70	0.643	0.88	0.959	0.80

Table 10: Per-class precision (P), recall (R), and macro-F1 for chunk-level English encoders.

A.5 Encoder HT Accuracy and Average MT Accuracy

Tables 11 and 12 report per-class HT recall and average MT recall (over translators) for the high- and low-resource datasets, respectively, complementing the balanced-accuracy numbers in Table 4 (averaging the two per row recovers the corresponding balanced accuracy).

A.6 Encoder Performance Across Languages and Datasets

Figure 1 visualizes the balanced-accuracy results of Table 4 across all languages, datasets, and input granularities.

A.7 Translator-Level Variability in MT Detectability

To better understand variability across translation systems, we provide box plots of MT detection performance computed over individual translators. Unlike aggregate metrics, these plots capture the distribution of detection accuracies across MT systems for each dataset and model.

High-resource languages. Figure 2 shows translator-level MT detection accuracies for high-resource datasets. We observe substantial variability across translators, particularly in the legislative domain (EU DGT), where interquartile ranges are wide for several models, indicating that detectability depends strongly on the specific MT system. Some settings (e.g., chunked models on OpenWHO) exhibit tighter distributions, suggesting more consistent behavior across translators.

Low-resource languages. Figure 3 presents the corresponding analysis for low-resource datasets. While overall detection is higher in many cases, variability across translators remains evident, particularly for Swahili. Some models (e.g., encoder-based classifiers) show consistently high detection across translators, while others exhibit wider spreads.

Language & Dataset	Model	HT recall	Avg. MT recall
English EU DGT	mDeBERTa	0.600	0.720
English EU DGT	mDeBERTa-chunk	0.714	0.900
English EU DGT	XLM-R	0.597	0.833
English EU DGT	XLM-R-chunk	0.737	0.584
English OpenWHO	mDeBERTa	0.367	0.688
English OpenWHO	mDeBERTa-chunk	0.670	0.802
English OpenWHO	XLM-R	0.333	0.691
English OpenWHO	XLM-R-chunk	0.755	0.965
English SMOL	mDeBERTa	0.405	0.763
English SMOL	mDeBERTa-chunk	0.812	0.702
English SMOL	XLM-R	0.357	0.724
English SMOL	XLM-R-chunk	0.643	0.959
Spanish EU DGT	mDeBERTa	0.634	0.666
Spanish EU DGT	mDeBERTa-chunk	0.662	0.718
Spanish EU DGT	XLM-R	0.632	0.648
Spanish EU DGT	XLM-R-chunk	0.774	0.766
Spanish OpenWHO	mDeBERTa	0.578	0.762
Spanish OpenWHO	mDeBERTa-chunk	0.738	0.822
Spanish OpenWHO	XLM-R	0.514	0.806
Spanish OpenWHO	XLM-R-chunk	0.712	0.888

Table 11: HT recall and average MT recall (over translators) for high-resource datasets.

Dataset	Model	HT recall	Avg. MT recall
Afrikaans FLORES+	mDeBERTa	0.624	0.716
Afrikaans FLORES+	mDeBERTa-chunk	0.962	0.978
Afrikaans FLORES+	XLM-R	0.612	0.668
Afrikaans FLORES+	XLM-R-chunk	0.854	0.886
Afrikaans SMOL	mDeBERTa	0.626	0.714
Afrikaans SMOL	mDeBERTa-chunk	0.932	0.948
Afrikaans SMOL	XLM-R	0.722	0.778
Afrikaans SMOL	XLM-R-chunk	0.784	0.816
Swahili FLORES+	mDeBERTa	0.616	0.704
Swahili FLORES+	mDeBERTa-chunk	0.712	0.748
Swahili FLORES+	XLM-R	0.602	0.678
Swahili FLORES+	XLM-R-chunk	0.732	0.768
Swahili SMOL	mDeBERTa	0.564	0.616
Swahili SMOL	mDeBERTa-chunk	0.602	0.638
Swahili SMOL	XLM-R	0.522	0.578
Swahili SMOL	XLM-R-chunk	0.814	0.866

Table 12: HT recall and average MT recall (over translators) for low-resource datasets.

A.8 Quality-Bin Analysis (COMET and BLEURT)

Figure 4 visualizes the COMET quality-bin detection rates of Table 7, and Figure 5 shows the corresponding BLEURT-based analysis.

A.9 Translator-Level Quality vs Detectability

Figure 6 plots average COMET score against MT detection rate at the translator level, across detectors, languages, and datasets, summarizing the quality–detectability relationship discussed in Section 5.

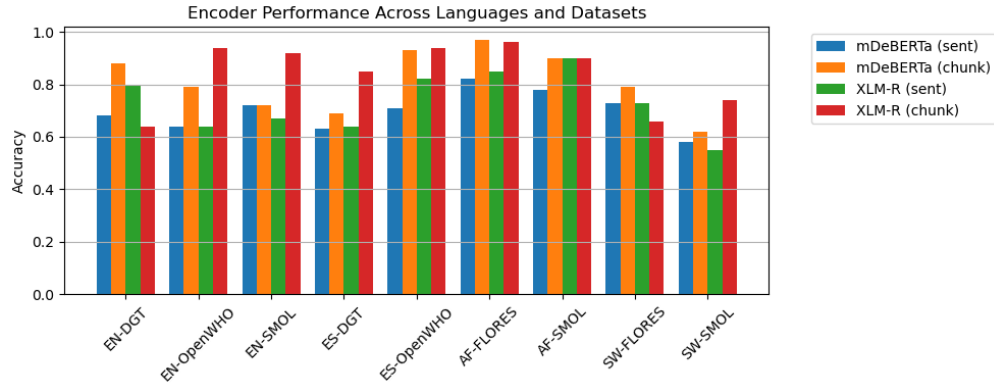


Figure 1: Encoder-based MT detection performance across languages and datasets. Results are shown for mDeBERTa and XLM-R under sentence-level and chunk-level (300-token) settings. High-resource results are reported for English and Spanish across EU DGT, OpenWHO, and SMOL; low-resource results for Afrikaans and Swahili across FLORES+ and SMOL. Chunking improves performance consistently for mDeBERTa, whereas the effect is more variable for XLM-R.

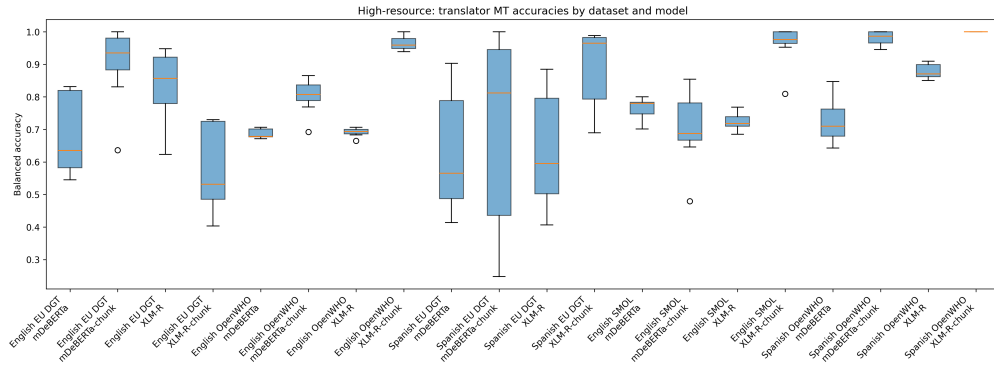


Figure 2: Translator-level MT detection accuracies for high-resource datasets. Each box is the distribution of detection performance across MT systems for a given dataset and model; wider boxes indicate greater variability across translators.

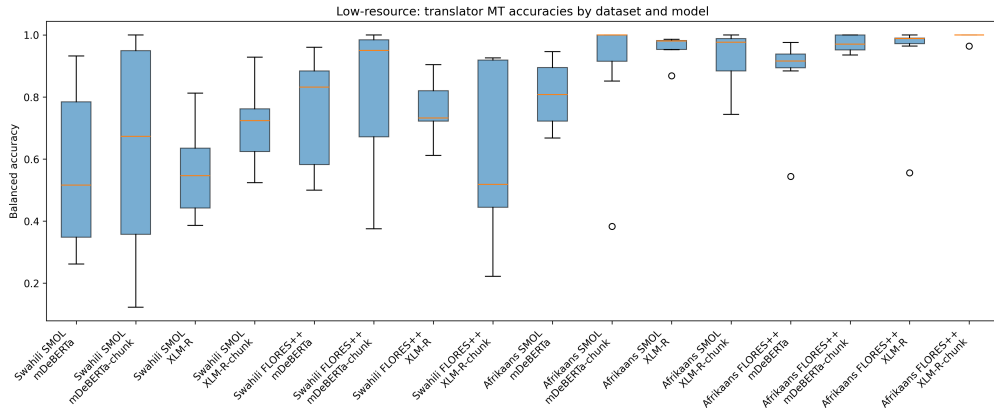


Figure 3: Translator-level MT detection accuracies for low-resource datasets. Variability across translators persists despite generally higher detectability.

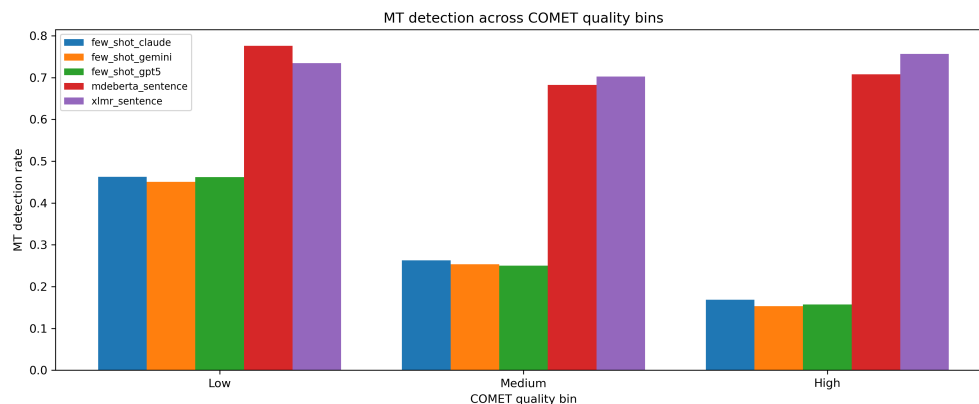


Figure 4: MT detection across COMET quality bins. For the few-shot LLMs, detection rates decrease as quality increases; sentence-level encoders remain relatively stable.

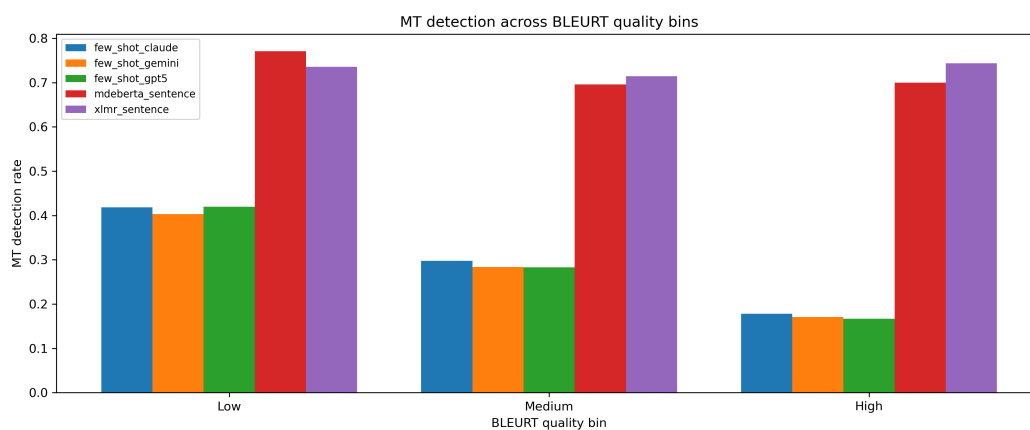


Figure 5: MT detection across BLEURT quality bins. Compared to COMET, the relationship between quality and detectability is weaker, although few-shot LLMs still detect lower-quality translations more often than higher-quality ones.

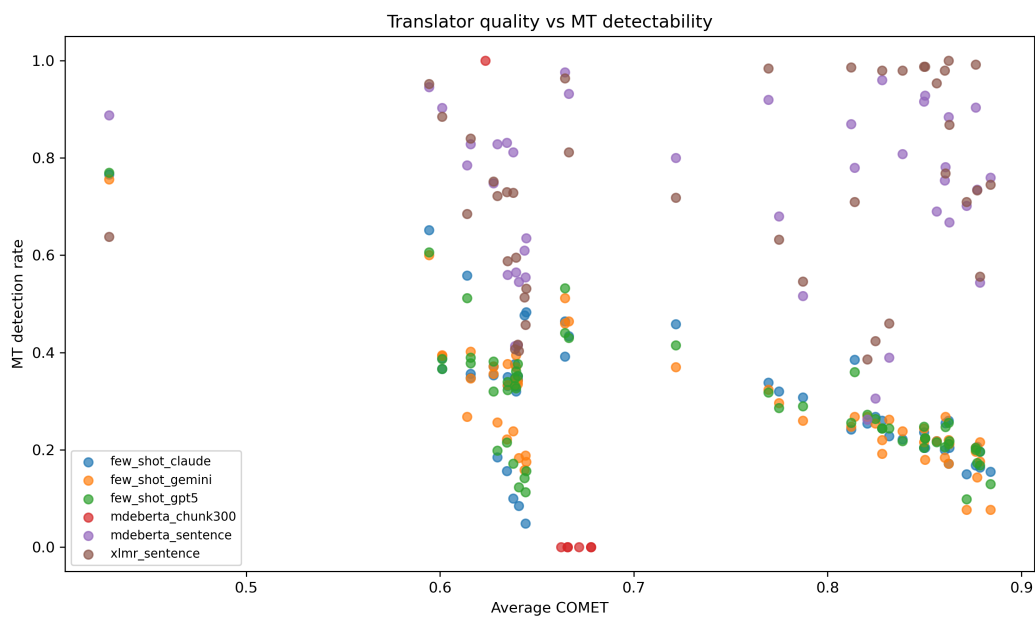


Figure 6: Translator-level relationship between average COMET score and MT detection rate across detectors, languages, and datasets. Few-shot LLMs show decreasing detection as translator quality increases, while encoder-based models exhibit weaker, less consistent dependence.