

KYX WMT 2026 CreoleMT System Description: A Contamination- and Orthography-Aware Evaluation for Réunion Creole to English

Yusuf Öztö

Department of Translation and Interpreting Studies

Boğaziçi University

yusufoztop33@gmail.com

Abstract

This paper presents the KYX submission to the WMT 2026 CreoleMT shared task for Réunion Creole to English (rcf-eng). It is mostly about how such a system should be evaluated. The pair is an extreme case: 191 training pairs, a 34-sentence development set, and no standard orthography. On a benchmark this small, a single chrF++ score is easy to over-read, so we pair our system with a contamination- and orthography-aware evaluation. The system continue-trains the official baseline adapter with QLoRA and raises the raw dev score by +13.3 chrF++. But a near-duplicate audit finds that six of the 34 dev sources are paraphrases of training sentences, which the exact-match check WMT26 mandates reports as 0% overlap. After decontamination the gain is +10.0. Our official test score then comes in 0.8 chrF++ below the decontaminated estimate and 5.1 below the raw one, but our margin over the baseline falls to +3.9, so even the adjusted gain was optimistic. A data-scaling ablation traces the gain to the provided real pairs, not to synthetic data. A source-spelling stress test then shows the advantage shrinks when the input is re-spelled into equally valid variants, though at this sample size the effect is not stable. We conclude that exact-match contamination checks are not enough on tiny benchmarks, and that scores like ours should be read with care. The same caution applies to any low-resource pair with only a few dozen dev sentences. We release the system, the audit, and all data.

1 Introduction

Extremely low-resource machine translation increasingly rests on very small benchmarks. For many endangered, indigenous, and creole languages, a development set of 30 to 50 sentences is close to all that exists, and at this scale evaluation is as much of a problem as modelling. Two artifacts can dominate a single automatic score. First, small train/dev overlaps that an exact-match check does

not catch can raise it, because even a few memorized paraphrases move the mean of a 34-sentence set. Second, when a language has no standard orthography, a character metric such as chrF++ partly rewards a system for reproducing the development set’s spelling rather than for translating well. Neither problem is visible in the number itself.

We study these effects on an extreme case, Réunion Creole to English (rcf-eng) in the WMT 2026 CreoleMT shared task (Robinson et al., 2026), which has 191 training pairs, a 34-sentence development set, and no standard orthography (the same word is written manzé, mangé, or manjé). We build a system for the pair, but our main point is about how to measure it: a single dev score does not, on its own, reflect the translation quality it appears to.

To that end we add three components to standard corpus chrF++ (§3): a near-duplicate audit for leakage that exact matching misses, decontaminated scoring for the inflation those near-duplicates cause, and a source-spelling stress test for orthography-driven overestimation. On rcf the audit flags six of the 34 dev sources as paraphrases of training sentences, which the exact-match statistic WMT26 mandates reports as 0% overlap, and re-scoring on the clean subset moves our system’s raw +13.3 chrF++ to +10.0. The official test scores, which arrived after these experiments, confirm that correction and then go past it: our test score comes in 0.8 chrF++ below the decontaminated estimate and 5.1 below the raw one, while our margin over the baseline is +3.9 rather than +10.0 (§4.4).

The system is a QLoRA continuation of the official baseline adapter, and we use it as a case study rather than as the main result. Its adjusted +10.0 chrF++ is stable across three seeds, but a data-scaling ablation shows the gain comes from the provided real pairs and not from synthetic data, and the spelling stress test suggests part of it is tied to the source’s surface form. That last effect is not stable across rulesets, so we report it as exploratory.

Contributions.

1. A contamination- and orthography-aware evaluation for extremely low-resource MT: three components, each tied to a concrete way a single chrF++ score misleads (§3).
2. Evidence that the exact-match contamination statistic WMT26 mandates is not enough on tiny benchmarks: it reports 0% overlap for rcf while six dev sources are paraphrase-level duplicates that move the score (§4).
3. A case-study system whose raw +13.3 chrF++ becomes an adjusted +10.0, with an ablation and a spelling stress test showing how little of the number reflects the language pair itself (§4, §5).
4. A held-out check on the protocol: on the official test set the decontaminated dev estimate predicts our own score to 0.8 chrF++ where the raw one is 5.1 optimistic, while the margin over the baseline falls from +10.0 to +3.9 (§4.4).

We release the system, the audit, and all perturbation data.

2 Task, Data, and System

Task. WMT 2026 CreoleMT, Subtask 1A, single unconstrained track (Robinson et al., 2026). We focus on a single direction, **rcf-eng**, evaluated with chrF++ (Popović, 2017).

Data. Réunion Creole is among the smallest pairs in the task (Table 1), which is drawn from the kreyòl-mt collection (Robinson et al., 2024). We train on 1,528 rows: all 191 provided pairs together with 1,337 synthetic rcf-eng pairs. A separate 150-sentence synthetic set is used only as an overfitting monitor, never to judge quality.

Synthetic data. The synthetic pairs are LLM back-translations of English sentences taken from OPUS-100 en-fr (Zhang et al., 2020). We stream that corpus, keep sentences of 3 to 14 words, and rank them by character-trigram Jaccard similarity against the English side of the 191 provided pairs, retaining the closest. Sampling English at random would have produced material from outside the narrow domain of the provided data; ranking against those pairs keeps the generated Creole in that domain. Claude Sonnet (Anthropic, 2026) then wrote a Réunion-Creole rendering of each sentence, prompted with the five nearest provided pairs as exemplars and with the French side of the OPUS sentence, rcf being French-lexified. Claude Opus

pair	train	dev	lexifier
hat-eng	665,033	646	French
mfe-eng	19,149	472	French
pcm-eng	5,545	311	English
rcf-eng	191	34	French

Table 1: Selected per-pair sizes (kreyòl-mt). rcf is an extreme case: a 34-sentence dev.

rated the 1,495 resulting pairs 4.63 out of 5.

That rating is the only quality signal available to us. Automatic quality estimation is not applicable here, since QE encoders have not been trained on rcf, so the adequacy of the synthetic data cannot be verified independently (§7). We then removed the synthetic pairs that the audit of §3 flags as near-duplicates of a dev item, which dropped 10, mostly variants of “What are you doing?”. Of the 1,485 that remain, 1,337 go into the training set and the other 148 make up all but two of the 150-sentence overfitting monitor. The ablation in §4 shows that the synthetic data contributes little to the decontaminated score in any case.

System. The baseline is google/gemma-4-E2B plus the official gemma4-creole-mt-2026-baseline-adapter (CreoleMT Shared Task, 2026), a generalist LoRA the organizers trained on public data across about 50 creoles. We *continue-train* this adapter with QLoRA, and select the winning epoch by chrF++ on the human dev, never by the synthetic eval-loss (which is anti-correlated with chrF++ here). Because we continue-train the provided adapter, the system and the baseline stay directly comparable, so any score difference between them comes from our training and not from a change of model.

Training details. We load the base model in 4-bit NF4 with double quantization and attach the official adapter with `is_trainable=True`, so training continues the organizers’ weights rather than initialising a new LoRA. The LoRA configuration is therefore inherited: $r=64$, $\alpha=128$, dropout 0.05, no bias, applied to all linear projections. We train for 8 epochs at learning rate 2×10^{-4} with `paged_adamw_8bit`, a linear schedule with no warmup and no weight decay, per-device batch 8 with gradient accumulation 2 (effective batch 16), and maximum sequence length 256. A checkpoint is saved every epoch and all 8 are scored; §4 reports the mean over seeds 42, 43, 44. Training

uses about 9.5 GB of VRAM. Decoding is greedy (do_sample=False) everywhere.

Metric. All scores are sacrebleu 2.6.0; chrF++ is CHRF(word_order=2). The signatures are
 nrefs:1|case:mixed|eff:yes|
 nc:6|nw:2|space:no|version:2.6.0
 nrefs:1|case:mixed|eff:yes|
 nc:6|nw:0|space:no|version:2.6.0
 nrefs:1|case:mixed|eff:no|tok:13a|
 smooth:exp|version:2.6.0
 for chrF++, chrF and BLEU respectively. Evaluation ran under Python 3.12 with transformers 5.14.1, peft 0.19.1 and torch 2.13.0.

3 The Evaluation Protocol

Standard practice reports one corpus chrF++. On a 34-sentence benchmark that number can be dominated by leakage or by orthographic variation, so our protocol adds one component for each failure mode:

problem with one chrF++ score	component
exact-match overlap misses paraphrases	(a) near-duplicate audit
near-duplicates bias the score	(b) clean-subset scoring
non-standard spelling inflates apparent quality	(c) spelling stress test

We never alter the official benchmark or the provided real data; we decontaminate only our own synthetic training data.

(a) Near-duplicate contamination audit. *Problem.* WMT26 asks teams to report train/dev overlap by *exact* string match on the *pair*, which flags only verbatim duplicates of a whole example, so paraphrase-level leakage passes straight through. *Solution.* We normalise both sides (NFKC, lowercase, strip punctuation, collapse whitespace) and score the two sides *separately*. For every dev item we find its most similar training source and, independently, its most similar training target, by a character-level difflib ratio. We flag the item if *either* side matches: source ratio ≥ 0.70 **or** target ratio ≥ 0.90 , and report this *near-duplicate* overlap alongside the exact-match statistic. Neither side is redundant: on rcf the source rule alone flags five items, the target rule alone flags two, and only one item (#24) is caught by both. Target-side leakage matters in its own right because chrF++ scores the English output, so a training example whose English matches a dev reference leaks the answer even when the two Creole sources are unrelated sentences (#12 in Table 2). The higher target threshold

reflects the fact that short English sentences collide by chance far more readily than Creole ones do. The ratio is deliberately conservative, since a semantic or embedding similarity would flag at least as many items, so our count is a lower bound, and the conclusion that exact matching misses paraphrases holds for any threshold below 1.0.

(b) Decontaminated (clean-subset) scoring.

Problem. Near-duplicates bias the mean, and on 34 sentences a handful of memorized items can dominate it. *Solution.* We report chrF++ on the full dev ($n=34$) and on the decontaminated subset ($n=28$) that drops the flagged items. The gap between the two is a direct read-out of near-duplicate-driven inflation, and the clean number is the one we treat as the contamination-adjusted estimate.

(c) Source-spelling stress test. *Problem.* With no standard orthography, a high chrF++ may reflect a model reproducing the dev’s particular spelling rather than translating well, and a character metric cannot tell the two apart. *Solution.* We re-spell each dev source into attested variants using the French-to-Creole correspondence rules described for rcf by Bollée (2013), and realised differently by the three competing orthographies in use on the island (Lékritir 77, KWZ, and Tangel). We then re-translate and measure how much of the system’s advantage over the baseline survives, against a random-noise control given the same number of edits (§5). Because both systems see the same perturbed input, we track the *change in advantage*, which within one ruleset does not depend on how harsh that ruleset is; comparing *across* rulesets does depend on it, and §5 shows this matters. This last component is exploratory: at $n=28$ its own conclusion is unstable, which is itself evidence that scores at this scale are fragile.

4 Applying the Protocol

4.1 Contamination audit and adjusted score

The statistic WMT26 mandates reports 0% overlap, and that is correct on its own terms: no training *pair* is a verbatim copy of a dev *pair*. Our audit reports six of 34 (indices 12, 21, 24, 27, 28, 32). The two statistics disagree because they measure different things. The mandated one compares raw strings and requires both sides of an example to match at once; ours normalises the text and scores each side on its own. Items #21 and #27 are character-identical to a training source after normalisation,

dev	flagged by	ratio	dev segment	nearest training segment
12	target (eng)	1.000	The sun has risen.	The sun has risen.
21	source (rcf)	1.000	La fam i ariv, pran lalé, kas son banann.	La fam i ariv, pran lalé, kas son banann.
24	both	0.765 / 1.000	komansman le mwá	ó kómansman du mwá
27	source (rcf)	1.000	Mwen lé fayfay.	Mwen lé fayfay.
28	source (rcf)	0.800	Moin la fine ariv la kaz.	Bann la i ariv la kaz.
32	source (rcf)	0.812	[...] li manz enn, li donn son syin sakin enn.	Li donn son syin sakin enn.

Table 2: The six flagged dev items, shown on the side that flagged them: rcf for source-side matches, English for target-side ones. #21 and #27 are byte-identical to a training source but their English differs (*I am tired.* against *I am a bit tired.*), so the pair is not a duplicate and the mandated check reports 0%.

but their English references differ from that training example’s English, so the pair is not a duplicate and the exact-match check passes them. In #27 the Creole source *Mwen lé fayfay.* appears in both, but the dev reference reads *I am tired.* where the training target reads *I am a bit tired.* Item #12 is the mirror image: its English is identical to a training target while the two Creole sources are entirely different sentences. Only #24 is close on both sides at once, and even there the Creole is a paraphrase rather than a copy. Table 2 lists all six.

Every one of the six traces to the real provided pairs rather than to our synthetic data. We therefore cannot drop them from training without making the baseline comparison unfair, since the baseline adapter was trained on the same public data; instead we drop them from scoring. On the full dev the system beats the baseline by +13.3 chrF++, and on the clean subset the advantage is +10.0 (Table 3). The 3.3-point gap between the two is the near-duplicate inflation the audit predicts, and it is visible in all three metrics we report (Table 3, bottom).

The six flagged items are also where the two systems differ most. Our system scores 72.8 chrF++ on them against 40.1 for the baseline, a gap of +32.7, while on the remaining 28 items the gap is +10.0. Six items scored three times better than the rest pull the 34-sentence mean up by 3.3 points on their own. Whether the underlying behaviour is memorisation or paraphrase generalisation cannot be settled at this sample size, and the two would be scored identically here; what the full-dev number cannot do either way is estimate quality on input the model has not seen. We therefore report both subsets.

4.2 Where the gain comes from

Because the data here is small enough to make the experiment cheap, we add training data in stages and score the best epoch by decontaminated chrF++

System	full ($n=34$)	clean ($n=28$)
Baseline (adapter)	36.26	35.49
Ours	49.60 $\pm$ 1.37	45.51 $\pm$ 2.01
Δ vs baseline	+13.3	+10.0
Merge (3-seed soup)	48.26	—

	full ($n=34$)		clean ($n=28$)	
metric	base	ours	base	ours
chrF++	35.98	49.47	35.24	45.20
chrF	35.68	49.36	35.01	44.98
BLEU	25.62	40.57	24.88	36.61

Table 3: rcf-eng. Top: chrF++, 3-seed mean $\pm$ std over the best epoch per seed. Bottom: the submitted system under three metrics; the advantage shrinks by 3.5 chrF++, 3.7 chrF and 3.2 BLEU on the clean subset, so the gap is not an artifact of the character metric. Each comparison is scored within one decoding environment.

at each stage (Table 4). Almost the entire decontaminated gain comes from the real pairs (+6.8 clean). Adding the synthetic pairs on top raises the clean score by only a further +1.1, which is not significant. So the clean-dev gain reflects the real pairs; the rest of the raw number is the contamination gap, not the extra data. The full-dev column stays between 47 and 48 once training starts while the clean column separates the two stages: on the raw benchmark the two trained systems look interchangeable, and only decontaminated scoring shows which addition did the work.

4.3 Statistical reporting

At $n=28$ a paired significance test has little power, so we treat intervals and effect sizes as more informative than p -values. The main effect size is the adjusted $\Delta = +10.0$ chrF++ (full-dev $\Delta = +13.3$). Across the three seeds the clean score is 45.5 ± 2.0 (mean $\pm$ std) against a baseline of 35.5, so every seed clears the baseline by a wide margin. A paired approximate-randomization test on the submitted system gives $p \approx 0.013$, but we do not rely on it,

training data	rows	full	clean	Δ prev
baseline (no training)	0	36.72	36.08	—
+ real pairs	188	47.82	42.84	+6.8
+ synthetic	1,528	47.53	43.97	+1.1

Table 4: Data-scaling ladder (single seed), chrF++ on the full and decontaminated dev; Δ prev compares clean scores against the row above. Neither step is individually significant on the clean subset ($p = 0.054$ and $p = 0.76$). Scored in a different environment from Table 3.

because with 28 items such a test is easily moved; the bootstrap intervals in §5 are the better guide.

4.4 Official test set

After the experiments above were complete, the organizers released scores on the held-out test set for rcf-eng, which contains 28 sentences (Table 5). Our system scores 44.38 chrF++ there.

Our own score transfers from the dev set, and it is the decontaminated estimate that transfers. That estimate sits 0.8 chrF++ above the test result where the raw dev score sits 5.1 above it, and chrF and BLEU behave the same way, with errors of 0.2 against 4.6 and 0.5 against 4.5. For the question of what our system scores on unseen input, dropping the six leaked items did most of the correcting that §3 claims for it.

The margin over the baseline does not transfer. The official Gemma baseline scores 40.49 chrF++ on this pair, so our advantage on the test set is +3.9 against the +10.0 we measure on the decontaminated dev subset. The difference comes from the baseline rather than from us. Our score moves by −0.8 between the two sets, while the baseline scores about five points higher on the test set than it does on our dev set. Dropping leaked items corrects the inflation that our own training introduced; it does nothing about two evaluation sets being differently hard for a system that has seen neither. We therefore read the +10.0 of §4 as an upper bound rather than an estimate, which is the more cautious of the two readings our protocol supports and the one the test set favours.

Some caution is due on the comparison itself. The organizers report chrF++ and BLEU but not chrF for their baselines, and their scores were produced in a different environment from our dev scores, where the baseline moves about 0.7 chrF++ (Table 3). The task’s second baseline, kreyòl-mt-pubtrain, scores 47.72 on the test set and is there-

	chrF++	chrF	BLEU
<i>our system, by evaluation set</i>			
raw dev ($n=34$)	49.47	49.36	40.57
decontaminated dev ($n=28$)	45.20	44.98	36.61
official test ($n=28$)	44.38	44.73	36.07
error of the raw estimate	+5.1	+4.6	+4.5
error of the clean estimate	+0.8	+0.2	+0.5
<i>official test set, task baselines</i>			
Gemma	40.49	—	30.89
kreyòl-mt-pubtrain	47.72	—	40.88

Table 5: The submitted system on the official test set. The decontaminated dev estimate predicts our own test score to within 0.8 chrF++ while the raw one is about 5 points optimistic, but our margin over the Gemma baseline is +3.9 rather than the +10.0 measured on the decontaminated dev. chrF is not reported for the baselines.

fore ahead of ours.

5 Exploratory Stress Test: Spelling Robustness

This test is exploratory rather than a main result. It asks whether the remaining +10.0 reflects translation quality or partly the dev set’s particular spelling, and what it shows most clearly is how unstable such a check is at this sample size.

The results are mixed. One of the three spelling rulesets we tried erases almost all of our advantage and the other two do not, and part of that disagreement comes from the rulesets perturbing the source by different amounts. The probe therefore shows that surface form matters here, without settling how much of the gain depends on it.

We re-spell each dev source into equally valid variants using the rcf spelling correspondences of [Bollée \(2013\)](#), re-translate, and measure how much of our advantage over the baseline survives. We compare three versions of the source: the original, the re-spelled one, and a control with the same number of random character edits, which changes the surface as much but follows no spelling rule. A gain that is real translation should survive all three. We repeat this with three rulesets (Table 6): v1, our main one; v2, a milder set; and v3, an accent-stripping set.

Under v1 the advantage almost vanishes under re-spelling, falling from +10.0 to +0.2 chrF++, while it stays at +8.9 under the random control. On its own this looks like the model leaning on the dev’s spelling rather than translating, and it is the only one of the three drops whose bootstrap

ruleset	edits		advantage	
	re-sp.	ctrl	re-sp.	ctrl
v1 (main)	9.9%	13.7%	+0.2	+8.9
v2 (mild)	4.9%	4.9%	+7.2	+3.1
v3 (accents)	7.4%	7.2%	+5.9	+3.0

Table 6: Spelling stress test on the decontaminated dev ($n=28$), where the unperturbed advantage is +10.0 chrF++. “edits” is the share of source characters changed. Drop in advantage under re-spelling, 95% bootstrap CI: v1 +9.8 [+1.3, +19.1], v2 +2.8 [−6.3, +12.1], v3 +4.0 [−4.9, +13.4].

interval excludes zero ([+1.3, +19.1]). But it does not replicate: under v2 and v3 re-spelling only lowers the advantage to +7.2 and +5.9, and it now hurts *less* than the random control rather than more.

Part of that disagreement is magnitude, not linguistics. v1 rewrites 9.9% of the source characters against v2’s 4.9% and v3’s 7.4%, and touches 30 of 34 sentences against 23 and 25, so it is simply a harsher perturbation and the ordering of the three drops follows the ordering of their edit rates. The within-ruleset control is the part that is properly matched (51 versus 51 edits for v2, 77 versus 75 for v3), and under v1 the control is if anything *harsher* than the re-spelling, at 142 edits against 102. v1’s result is therefore not an artefact of its spelling perturbation being the larger of the two.

The two systems respond differently to the same perturbation. Under v1 the baseline loses 1.1 chrF++ to re-spelling while our system loses 10.9; under the random control they lose 12.1 and 13.2. The baseline is thus insensitive to re-spelling specifically, not to surface change in general, whereas our system is not. Continued training on 191 sentences written in one orthography is the obvious candidate explanation, but we cannot separate it from the other differences between the two systems here.

The probe therefore suggests that part of the gain is tied to surface form, but cannot say how much. At 28 sentences even a purpose-built check is unstable, and comparing across rulesets confounds the severity of a perturbation with its content. We report it as supporting, exploratory evidence that a single score here should not be read at face value.

6 Related Work

Contamination in evaluation. Benchmark leakage is a recognized confound, systematized by the CONDA shared task (Sainz et al., 2024); for MT, Kocyyigit et al. (2025) show that source–target con-

tamination inflates scores, motivating the overlap reporting WMT26 mandates. Near-duplicate leakage of the kind we target is the problem addressed by MinHash deduplication (Lee et al., 2022). The WMT25 Creole task made a related point empirically (Robinson et al., 2025): contamination was serious enough to require the organizers’ own decontaminated re-evaluation, and EdinHelsOW reported gains that held only on the decontaminated test set (Rowe et al., 2025). That precedent is the closest existing answer to whether other creole MT tasks would benefit from filtering of this kind: on those pairs decontamination already changed which systems looked better, but it was applied after the fact and only for exact overlap. Our result indicates that the same correction is needed before scores are reported, and at a looser matching threshold than exact. How much leakage a paraphrase-level audit would find on those larger pairs remains open.

Orthographic variation and robustness. chrF++ (Popović, 2015, 2017) is character-based, so it is sensitive to spelling: Aepli et al. (2023) show that non-standard spelling degrades metric reliability on dialects without a standard orthography. Robustness of NMT to input noise is well studied (Belinkov and Bisk, 2018), but for creoles no prior work probes the model’s own spelling robustness as part of evaluation. Our stress test does, and finds it unstable at $n=28$.

7 Limitations

The evidence base is 34 sentences (28 decontaminated), so all point estimates carry wide intervals and every selection step is noisy: the best epoch varies across seeds, and chrF++ swings up to about 5 points across epochs within a seed (42.96 to 49.47 on the full dev). Reporting over three seeds reduces but cannot remove this fragility, and the stress test (§5) inherits it; cross-validation would give a fuller picture of the variance and we leave it to future work.

The audit thresholds (0.70 and 0.90) are heuristic, though our claim is threshold-independent for any cutoff below 1.0; a looser rule flags eight items instead of six, and the direction of every conclusion is unchanged. We evaluate a single direction (rcf-eng) and do not report into-Creole results, where chrF++ is unreliable for the same orthographic reasons. The single-direction scope is the main limitation of this work: we show that the mandated exact-match statistic misses paraphrase-level

leakage on rcf, but not that this holds for the task as a whole.

Moving the spelling component to a second pair is harder than moving the audit, because most of the collection does not have the property that component tests. Haitian, Mauritian and Seychellois Creole have had codified orthographies since 1979, 2011 and 1981 respectively, and Guianese Creole is written in the semi-official GEREK system; rcf is written in at least three competing ones and has no agreed standard. Louisiana Creole is the closest parallel in the collection, with several orthographies in concurrent use and a 92-sentence development set, and it is where we would take the spelling test next. That requires a documented correspondence set for Louisiana Creole rather than a reuse of ours, so we do not report it here. COMET-style metrics are omitted as primary because their encoders have not seen rcf, and for the same reason we could not apply quality estimation to the synthetic training data, whose adequacy therefore rests on an LLM judgement alone.

We also do not test whether a long-context model prompted with the entire 191-pair training corpus, rather than fine-tuned on it, would score differently. Such a system retrieves from a visible corpus instead of memorising into weights, so it might show a different gap between the raw and the decontaminated score, and the comparison would isolate what fine-tuning contributes over in-context access to the same data. We leave that to future work. Official test-set scores are reported in §4.4. They show that the decontaminated dev figure, which is the estimate this paper argues for, still overstated the margin over the baseline by roughly a factor of three, so the protocol narrows the error on a benchmark this small without removing it.

8 Conclusion

Benchmarks of a few dozen sentences are now the norm in extremely low-resource MT, and at that scale a single chrF++ score should be read with care rather than at face value. We proposed a contamination- and orthography-aware evaluation, a near-duplicate audit, decontaminated scoring, and a source-spelling stress test, and applied it to WMT 2026 Réunion Creole to English. The audit exposes paraphrase-level dev/train leakage that the exact-match statistic WMT26 mandates reports as 0%, which moves our system’s raw +13.3 chrF++ to an adjusted +10.0, and the ablation and stress test

show how little of even that number reflects the language pair rather than the data and its spelling. The gap between the two numbers is a property of the benchmark, not of the system: it is produced by six items that any exact-match check passes, and it survives a change of metric. The official test scores arrived after the analysis was complete. They put our system 0.8 chrF++ below the decontaminated estimate and 5.1 below the raw one, and they put our margin over the baseline at +3.9 rather than +10.0, so the adjusted dev figure was the closer of the two and still too high. We release the system, the audit, and all data.

References

- Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. [A benchmark for evaluating machine translation metrics on dialects without standard orthography](#). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.
- Anthropic. 2026. [Claude](#). Large language model.
- Yonatan Belinkov and Yonatan Bisk. 2018. [Synthetic and natural noise both break neural machine translation](#). In *International Conference on Learning Representations (ICLR)*.
- Annegret Bollée. 2013. [Reunion creole](#). In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *The Survey of Pidgin and Creole Languages, Volume II: Portuguese-based, Spanish-based, and French-based Languages*, pages 241–249. Oxford University Press, Oxford.
- CreoleMT Shared Task. 2026. [CreoleMT 2026 baseline adapter for google/gemma-4-E2B](#). Hugging Face model repository.
- Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. [Overestimation in LLM evaluation: A controlled large-scale study on data contamination’s impact on machine translation](#). In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pages 31105–31132.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation (WMT)*.
- Nathaniel Robinson, Ruth-Ann Armstrong, Claire Bizon Monroc, Rasul Dent, Pranav Gupta, Raj Dabre, Andre Coy, and Kenton Murray. 2026. Findings of the second shared task for creole language machine translation, at WMT 2026. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Kenton Murray, Raj Dabre, Andre Coy, and Heather Lent. 2025. [Findings of the first shared task for creole language machine translation at WMT25](#). In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 520–531.
- Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Ones, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Dean Stutzman, Bismarck Bamfo Odoo, Sanjeev Khudanpur, Stephen D. Richardson, and Kenton Murray. 2024. [Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African creole languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3083–3110.
- Jacqueline Rowe, Ona de Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch, and Yves Scherrer. 2025. [Edinhsow WMT 2025 CreoleMT system description: Improving lusophone creole translation through data augmentation, model merging and LLM post-editing](#). In *Proceedings of the Tenth Conference on Machine Translation (WMT)*.
- Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D’Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, and 9 others. 2024. [Data contamination report from the 2024 CONDA shared task](#). In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 41–56.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. [Improving massively multilingual neural machine translation and zero-shot translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1628–1639.