

ChapsVision (CHV) at WMT 2026 CreoleMT: Shifting Creole MT from Train-Time to Test-Time Compute for Haitian Creole

Louis Jourdain

ChapsVision

Université Paris-Saclay

ljourdain@chapsvision.com

Aurélié Pradelles

ChapsVision

apradelles@chapsvision.com

Nasredine Semmar

Université Paris-Saclay, CEA, List

Palaiseau, France

nasredine.semmar@cea.fr

Abstract

We describe a training-free system for the WMT26 Creole MT shared task, targeting Haitian Creole in four directions (eng↔hat, fra↔hat). Rather than fine-tune, we wrap a single fixed 8B instruction model (qwen3-8b) in a language-agnostic *harness* that retrieves lexical, parallel-data and grammatical evidence and injects it directly into the translation prompt. On FLORES+ devtest the harness lifts the base model by up to +14 chrF++ and clears the shared task’s primary baseline in all four directions, though a purpose-fine-tuned system (kreyòl-MT) and frontier LLM still lead the mean. On the task’s blind test set that ordering reverses: we finish below both baselines, while the harness gain itself holds, staying large into Creole and small out of it. Two findings frame the paper. First, injected knowledge behaves as a reasoning *equalizer*: the same evidence that lifts a weak open model leaves a strong frontier model flat, so grounding substitutes for a capability the model lacks rather than amplifying one it has. Second, how evidence is delivered dominates: forcing it into the prompt beats letting the model self-serve the same tools in an agentic (ReAct) loop, which costs about 6× the tokens.

1 Introduction

Machine Translation naturally followed (or initiated) new model architectures, each one significantly improving translation quality, but the benefits remain highly language-dependent. While MT systems approach human production for high-resource language pairs with massive training corpora available (NLLB Team et al., 2022), such resources are lacking for low-resource languages and domains. This is especially true for Creole languages which, while they share lexical similarities with high-resource languages, display specific syntactic features (DeGraff, 2007), a distinct referential universe, and uneven coverage (Lent et al., 2024).

This discrepancy suggests that training-based approaches alone seem unable to bridge the quality gap between high- and low-resource languages, and that other methods should be explored: even the largest dedicated Creole corpus to date (Robinson et al., 2024) leaves most of its 41 Creoles with too little aligned text to train on.

The pivot of the MT community towards LLMs, which acquired natural translation abilities during post-training (Wei et al., 2022), enables the exploration of such approaches. Contrary to former specialized MT models, an LLM can be prompted with external information — vocabulary (Ghazvininejad et al., 2023), example pairs (Mu et al., 2023), or grammar excerpts (Tanzer et al., 2024; Marmonier et al., 2025) — that is proven to enhance translation quality. Since 2023, LLMs have undergone several major innovations: the shift from train-time to test-time compute integrated the benefits of Chain-of-Thought prompting into the model’s behaviour; and tool-calling capabilities give the model autonomous access to probe external data sources to accomplish its task (ReAct; Yao et al., 2023). The application of these reasoning and agentic capabilities is still being explored by the MT community (Briva-Iglesias, 2025), mainly on well-established language pairs and at a token cost that has so far outrun its measured benefit (Briva-Iglesias and Dogru, 2025). However, these approaches offer an interesting opportunity for languages for which aligned corpora are scarce yet linguistic documentation is satisfactory. In designing such systems, the researcher’s effort shifts from optimizing the learning phase to orchestrating the model’s prompt content and task decomposition (*context engineering*; Mei et al., 2025).

The WMT26 Creole MT shared task (Robinson et al., 2026) provides a promising testbed for this shift. We target Haitian Creole, in the four directions eng↔hat and fra↔hat, evaluated with

chrF++ (Popović, 2017) using the FLORES+ benchmark (Open Language Data Initiative, 2024) to calibrate the model’s harness. Haitian Creole is emblematic of the low-resource condition: it counts millions of speakers yet offers only a thin band of aligned text. Under such scarcity, training-based approaches face a ceiling: the parallel data required to close the quality gap simply does not exist. That ceiling motivates a complementary route: rather than have the model learn a language from data it does not have, we inject at inference time the lexical and grammatical knowledge that linguistic documentation does provide, and let a general-purpose model exploit it. The question is then not how much parallel data would close the gap, but how cheaply documentation can stand in for it. We explore this through three questions.

Research questions.

RQ1 Does the shift from train-time to test-time compute pay off? Can a single unmodified open model fed with retrieved evidence become competitive with purpose-trained systems on a genuinely low-resource pair, and at what inference cost?

RQ2 How should evidence reach the model? Does letting the model self-serve tools and reason over them in an agentic loop help, compared to forcing the same evidence directly into the prompt?

RQ3 When does inference-time knowledge injection help? Is its benefit a fixed property of the resources, or does it depend on what the base model already knows?

Contributions. Our main contributions are sketches of answers to these questions, for Haitian Creole at least. We present:

- **A competitive training-free harness (RQ1).** A fine-tuning-free harness (a fixed 8B model wrapped with retrieved lexical, parallel and grammatical evidence) matches the shared task’s purpose-trained primary baseline in all four directions on our dev set, though a fine-tuned system (kreyòl-MT) still leads the mean, frontier models lead overall, and on the blind test set we finish below both baselines (Appendix F). This is the first systematic tool-augmented, training-free system built for the WMT Creole task (Robinson et al., 2025).

The closest prior work is Pei et al. (2025), who treat the prompt as a modular space and ablate four knowledge types on Manchu; on the agentic side, prior multi-agent MT chains specialized LLM calls without autonomous tool use (Kim, 2025). Appendix A surveys both lines in full.

- **Delivery dominates over agency (RQ2).** Holding the tool set fixed, forcing evidence into the prompt beats letting the model self-serve the same tools in an agentic (ReAct) loop in every direction, and costs about $6\times$ fewer tokens; richer multi-agent orchestration is neutral-to-negative.
- **Tools as a reasoning equalizer, not an amplifier (RQ3).** The same evidence bundle that lifts a weak open model by up to +14 chrF++ leaves a strong frontier model flat (≈ 0). Grounding seems not to always scale ability up but to substitute for an ability the model lacks. Therefore the benefit of inference-time injection is related to the base model’s own literacy.

2 Methodology

2.1 Target language and directions

We focus on Haitian Creole (*Kreyòl*), a French-lexifier, isolating language with essentially no inflectional morphology, SVO order, and preverbal tense–mood–aspect markers with roughly 12 million speakers and co-official status with French since the 1987 constitution. It is comparatively well documented: a standardized orthography, developed by the Institut Pédagogique National (IPN) and given official status with the 1979 education reform, then fixed by the communiqué of the Secrétairerie d’État à l’Éducation Nationale of 22 January 1980 (Bernard, 1980; Faraclos et al., 2010); reference dictionaries (Freeman, 2010) and grammars (DeGraff, 2007); and very much alive online, so monolingual and reference material is comparatively plentiful in LLM pre-training. We evaluate on the four directions, both orders of each pair: *eng*↔*hat* and *fra*↔*hat*.

2.2 A language-agnostic machine-translation harness

The goal of our study is not limited to building a translation system but to reach an understanding of which knowledge helps an LLM for a low-resource

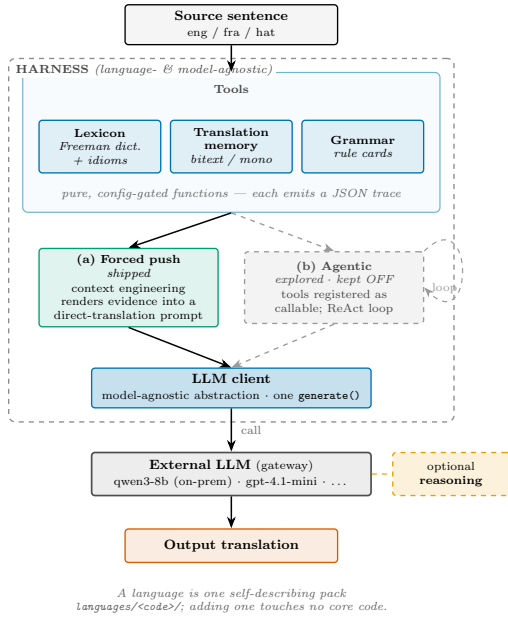


Figure 1: **Architecture of the harness**

language. We take inspiration from the recent notion of *context* and *harness engineering* — building an engineered scaffold around an LLM that orchestrates the content fed to the model — to design a modular system around the model, in which each form of injected knowledge is an independently switchable component (Figure 1).

This architecture lets us compare system configurations easily, and is model-, tool- and language-agnostic. Contrary to much work on context engineering (Mei et al., 2025), each knowledge source is treated as a harness tool: a function, parameterised by configuration, that emits a JSON trace. This build-once discipline makes the ablations homogeneous and every run replayable, and each tool carries options (feature flags) that support rigorous ablation; Appendix B lists every flag with its shipped value.

The harness is also language-agnostic by construction and offers a standardised schema for dictionary lemmas and translation-memory examples, letting the user load new data.

The tools can either be activated by the LLM in an agentic loop, or called by default and integrated into the prompt, a mode we call *forced push* which we find more beneficial (§3.2).

2.3 A methodic harness-improvement protocol

Our development set is the FLORES+ devtest split (Open Language Data Initiative, 2024) (about

1,000 sentences per direction),¹ not the Kreyòl-MT dev split the task distributes: the task’s own test sets are blind and source-only, and that dev split is largely not made of sentences — 11.8% of its targets are under four tokens and our ingestion gate flags 4.3% of its pairs as defective, against 0% and 0.1% for FLORES+, which is also the set on which we score every external baseline.² This set serves as our dev set: we iterate versions of the harness changing one parameter at a time (A/B testing), for instance tuning a tool hyperparameter. To assess a lever we calibrate on the first 200 sentences of a direction, then confirm the winners on the full devtest. Significance is established by paired bootstrap resampling (Koehn, 2004) on matched indices: we promote a lever only when it clears $p < 0.05$ against the arm it replaces. This guards against overfitting a decision to a 200-sentence slice, which we have repeatedly watched happen. We also submitted to the task’s blind challenge test, whose official scores the organizers released after the deadline; they are reported in Appendix F.

The official metric is chrF++ (Popović, 2017), computed with sacrebleu (Post, 2018) (character order 6, word order 2). The harness also emits chrF and BLEU for cross-checking.

We performed our calibration on *qwen3-8b*, a small open-weight model hosted on premises. We measure against two external systems: the shared task’s primary Gemma baseline (Gemma-4-E2B, $\approx 2B$ effective parameters, with a Creole-MT LoRA adapter) (Gemma Team, Google DeepMind, 2025), the score to beat,³ and kreyòl-MT (Robinson et al., 2024), the mBART-50 (Tang et al., 2020) Creole fine-tune whose public-data variant is the task’s secondary baseline.

¹Two devtest rows have no correct Haitian text — idx 277’s Haitian slot duplicates idx 37’s sentence, and idx 764’s holds a sentence belonging to no English row — and are dropped, so every number in this paper, ours and every baseline’s, is scored on the same $n=1010$ index set.

²The gate detects residual markup, non-linguistic strings, untranslated identical pairs, extreme length ratios, mismatched numbers on short segments, and adult or crude language. The wider pool is worse than its dev sample: 33% of its targets are under four tokens.

³Gemma-4-E2B *base* (no adapter) scores above its own LoRA-adapted variant on ht→en (Table 1) because it largely ignores the requested target language and answers in English: 99% of its ht→fr output is English and into Creole it often copies the English source verbatim. ht→en is the single direction in which that default behaviour is the correct one. The adapter’s main contribution is target-language compliance.

2.4 Harness tooling

Three knowledge tools were integrated into the harness, each assembling deterministic evidence for the generation step. Worked input/output for each tool is shown in Appendix C.

2.4.1 Lexicon tool

The most direct symptom of low-resource translation is a vocabulary gap: the base model simply does not know enough Creole words. Our first module attacks this gap in the spirit of dictionary-prompted MT (*DiPMT*) (Ghazvininejad et al., 2023): For Haitian a large reference lexicon, *Freeman* (6th ed., 2010, University of Kansas; CC BY-NC-ND) (Freeman, 2010), 78,843 bilingual entries was available (full inventory in Appendix D).

We work from a scanned OCR dictionary published on the University of Kansas repository, and had to curate and organize the lemmas and fix OCR mistakes (especially on accented *o* and *e*). The lexicon tool takes a list of words to look up and returns, for each, a structured entry storing its lexical information. We refined it in several directions.

(i) *Incorrect function-word glosses*. Mis-parsed dictionary entries injected wrong senses for high-frequency function words, for instance *it* matched *piki*, which the model dutifully copied; we quarantined these at the data layer. (ii) *Hidden common lemmas*. OCR had made many high-frequency headwords unreachable through POS-tag bleed, glued multi-entry lines, and de-columnization losses; we recovered several thousand entries. (iii) *Noisy entries*. Because the raw gloss field is a verbose dump (POS tag, numbered senses, example tails), a deterministic cleaner extracts the leading Creole rendering of each sense in dictionary order. (iv) *Inflection*. A source-morphology lemmatizer was used on the source language to improve lemma match. Finally we also incorporated *idioms*: Freeman’s $\approx 18\text{k}$ idiom sub-entries are matched by exact phrase or gloss overlap (into Creole), so a fixed expression surfaces as a whole rather than being calqued word by word.

A recurring observation which we return to in §3.5 is that these are correctness wins that leave the aggregate score close to unmoved: cleaner evidence does not rescue a model whose bottleneck is fluency. We A/B each knob; the per-lever verdicts are in §3.4.

2.4.2 Translation memory

Our second module reuses parallel data at inference in the spirit of translation-memory prompting (Mu et al., 2023). The tool takes the source sentence, finds its nearest neighbours in a corpus of sentence pairs aligned across the two languages and injects the retrieved sentence pairs into the prompt. We build it from the Kreyòl-MT corpus on Hugging Face (on the order of 665k English–Haitian pairs).

Concretely: every indexed corpus is embedded offline with *bge-m3* (Chen et al., 2024a) and the index persisted, so at inference we embed only the query sentence (one call) and run a cosine search against the stored vectors. The query is matched monolingually against the source side of each bitext. Candidates below `min_sim` are discarded, survivors are de-duplicated and noise-filtered, and the top k that remain are injected as source–target pairs. Corpora that are evaluation sets are excluded.

Retrieval varies along two axes: what the corpus offers, and how we use the hit. A bilingual corpus gives sentence pairs and is matched against its source column, in the language of the query; a monolingual corpus gives target-language sentences only and must be matched across languages.

The retrieved pairs, shown as loose exemplars, lift both into-Creole directions, with a fixed fallback of professional FLORES-dev pairs when nothing clears the bar, which happens since the parallel data is religious, health and governmental where FLORES is encyclopedic and news, so a novel composition has no near-twin to copy, and switching from lexical overlap to dense embeddings does not change that. The monolingual channel injects fluent Creole sentences as a fluency prior and helps marginally, into Creole only. What decides the outcome is therefore not the matching heuristic but how the model exploits it.

Two hyperparameters govern retrieval, and both were fixed under the calibrate-then-confirm protocol. The number of exemplars per channel is $k=3$: fewer under-primes, and the gain plateaus beyond three. The cosine-similarity floor is `min_sim=0.55`; a looser 0.40 won on the 200-sentence calibration slice but lost on full devtest. Per-regime deltas are in §3.4.

We also incorporated 5 fixed parallel sentences (not tailored from the input) in the system prompt as few shot examples for the translation task as it enhanced translation quality overall.

2.4.3 Grammar tool

Our third module targets the morphosyntax a dictionary cannot: preverbal tense–mood–aspect markers (*te*, *pral*, *ap*), postposed determiners, negation *pa*, and *se* clefts. We build it from a reference grammar (DeGraff, 2007): an ingestion pass distils it using an LLM into 115 rule cards (phenomenon, statement, glossed examples), indexed with *bge-m3* embeddings. The main difficulty is selecting the cards relevant to an input sentence; we therefore use feature-aware retrieval to flag the marked constructions and boost the matching cards.

The payoff is marginal and direction-dependent: that grammar helps least, despite a retrieval layer built specifically to make rule selection work, echoes Pei et al.’s Manchu finding (Pei et al., 2025) that explicit grammar adds little once the lexicon is in place. Per-direction Δ chrF++ verdicts are in §3.4.

2.5 Comparing agentic delivery to forced knowledge injection

One of the goals of this work was to determine whether an agentic approach helps compared to forced knowledge injection. We hoped to leverage the model’s reasoning abilities to act as a professional translator that analyses the sentence’s structure, fetches the translation of unknown words, and picks the best rendering.

In the *agentic* regime the model drives a ReAct loop: it emits tool calls we execute and integrate until the final translation is uttered (or a step cap forces one). In the *forced-push* regime the same lexical, parallel-data and grammar evidence is rendered up front into a plain direct-translation prompt for a single inference, and no tool is exposed. If the underlying model allows, reasoning can be freely activated in either mode. The two regimes are cleanly separable: a channel can inject its evidence while its callable tool stays off, which is what lets us hold the evidence fixed and vary only the delivery. Which of the two we ship is itself one of our results (§3.2).

3 Results

3.1 A training-free harness significantly lifts the score of an off-the-shelf 8B model

Injecting knowledge at inference time moves the base model far enough to matter. On our dev set, the harness lifts *qwen3-8b* from well below the WMT26 primary baseline (Gemma-4-E2B with the

Creole-MT LoRA adapter) to above it in three of the four directions. On *ht→fr* it lands just short (50.6 against the baseline’s 51.0) and reasoning is what settles it, at 53.6. Gating reasoning to that one direction — the shipped configuration — therefore beats the baseline in all four at once and on the mean (47.8 vs. 45.7 chrF++; Figure 2, Table 1). A frugal, training-free scaffold on an unmodified 8B model is thus competitive with a purpose-fine-tuned 2B system on this set with no training at all.

The gain is far from uniform and concentrates on the low-resource side. The bare model is weakest into Creole, at 28.5 (*en→ht*) and 30.6 (*fr→ht*), nine to twelve points under the baseline, and that is where the harness does most of its work: it adds +14.1 and +12.1 there, against +3.6 and +3.9 out of Creole. Those deficits become gains over Gemma of +2.3 and +2.7. Each decisive into-Creole lever is significant at $p < 0.001$ under a paired bootstrap. We show this for Haitian Creole and the models we tested; that the into-vs-out gap reflects resource geometry rather than any one dictionary is a hypothesis, not a cross-language result.

Beating the primary baseline does not, however, solve Creole MT: fine-tuned Kreyòl-MT still leads the mean (50.1) and the frontier model we tested, *gpt-4.1-mini*, sits well ahead (58.4), a residual we argue is base-model fluency.

That dev-set result, however, does not survive the change of domain. On the task’s blind test set our submitted system scores below both organizer baselines in all four directions, by 2.6 to 7.3 chrF++, and behind every other team that entered the same direction; every system drops sharply between the two sets and ours drops the most (Appendix F). What does transfer is the mechanism: on the blind set the harness still adds +8.2 and +5.4 chrF++ into Creole against +1.7 and +0.5 out of it, and reasoning still helps only *ht→fr*. We therefore claim a dev-set result against the primary baseline and a blind-set confirmation of where grounding pays, not a shared-task win. Appendix E gives side-by-side outputs of all five systems on sample sentences.

3.2 Agentic delivery and reasoning do not push knowledge injection further

Agentic delivery does not help in the harness’s current configuration. Holding the tool set fixed at the best configuration (few-shot, gloss lemmatizer, anti-code-switching, token budget), we vary only how evidence reaches the model (forced into

System	Params	Into Creole		From Creole		Mean
		en→ht	fr→ht	ht→en	ht→fr	
<i>Baselines</i>						
Gemma-4-E2B base (no adapter)	~2B	31.7	18.1	55.5	20.6	31.5
+ Creole-MT LoRA (WMT26 primary)	~2B	40.4	40.0	51.5	51.0	45.7
Kreyòl-MT (mBART; WMT26 secondary)	~0.6B	49.8	47.2	50.9	52.3	50.1
<i>Our system</i> (qwen3-8b)						
base (no augmentation)	8B	28.5	30.6	48.7	46.7	38.6
+ harness	8B	42.7	42.7	52.3	50.6	47.0
+ harness + reasoning	8B	41.9	41.4	52.5	53.6	47.4
shipped (reasoning ht→fr only) [†]	8B	42.7	42.7	52.3	53.6	47.8
<i>Frontier reference</i> (gpt-4.1-mini, not eligible for bold)						
direct (no tools)	—	53.4	53.0	63.7	63.7	58.4

Bold = best per column (frontier reference excluded as an upper bound); **green** = beats the WMT26 primary baseline.

[†] Reasoning is kept for ht→fr only, the one direction where it is clearly beneficial (+3.03 chrF++, curing the Creole-passthrough collapse of that French-lexified direction). Elsewhere it costs rather than pays (§3.4).

Table 1: **chrF++ on FLORES+ devtest, four directions** (idx-matched $n=1010$). The *shipped* row takes, per direction, the better of the reasoning-off and reasoning-on variants above it.[†]

the prompt vs. agentic ReAct loop) and whether reasoning is on, across the first-200 slice of all four directions (see a comparison in Appendix G.1–G.2). Forced push wins in every one of the eight cells. Crucially the margin tracks how much the target side needs the injected knowledge: into Creole, agentic trails forced by about eight chrF++ (en→ht 34.8 vs. 42.9, fr→ht 34.7 vs. 43.1); out of Creole, where the model already renders fluent English or French, the gap collapses to two points or less (ht→en 50.7 vs. 52.8, ht→fr 50.5 vs. 50.9). Native reasoning does not rescue agency (Figure 3).

The reason is behavioural: the weak model spends its agentic step budget on one-word-per-turn lookups and sometimes falls into a repetitive tool-calling loop, rather than composing the evidence into a translation.

Beyond the ReAct loop we tried the richer multi-agent orchestration patterns recent work proposes (Kim, 2025), all of which ask the weak model to reason more about the task: decomposing the knowledge push into specialist consultants (each emitting typed evidence for a synthesizer, to sharpen word-sense disambiguation and gloss use), selecting among several candidate drafts with a reference-free judge, and revising its own output in a second pass. Every one was neutral-to-negative (Table 2). Specialist decomposition is the worst: the model’s LLM-generated evidence invents wrong senses that the synthesizer then calques, and deterministic dictionary gloss beats it by up to eight chrF++. Self-revision is dead, and ensemble selection is at best neutral. The pattern is consistent: the informed weak model is a better

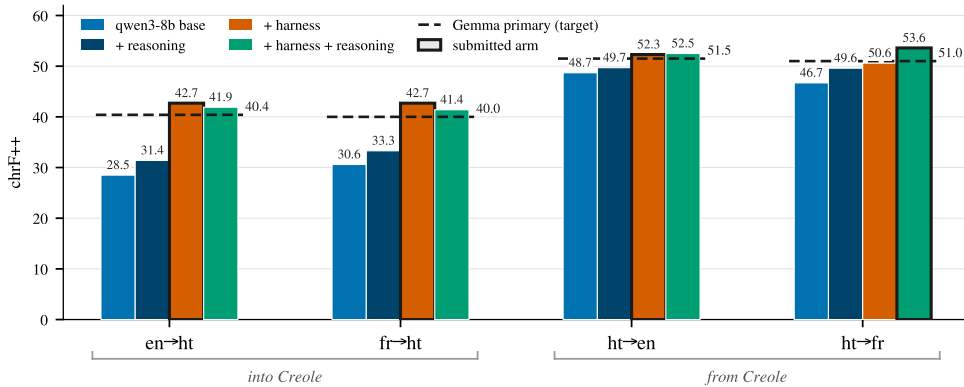


Figure 2: **Per-direction chrF++ against the Gemma baseline (dashed); submitted arm outlined.** Reasoning alone lifts every direction, but on top of the harness it pays only for ht→fr: the two are substitutive, not additive.

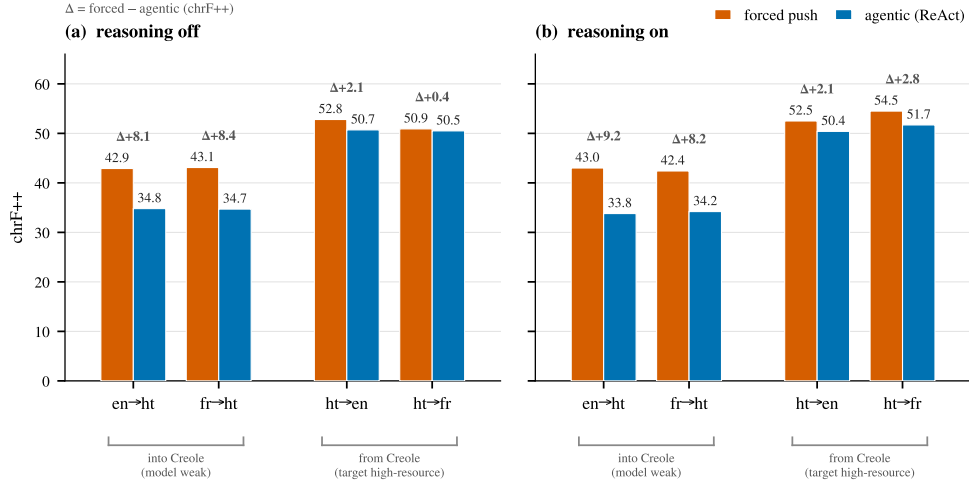


Figure 3: chrF++ by direction, forced-push vs. agentic delivery of the same tools (first-200, qwen3-8b): reasoning off (a), on (b); Δ = forced – agentic.

Configuration	Mechanism	Δ chrF++
Specialist decomposition	typed-evidence consultants → synthesizer	−8.4
Sense-select sub-agent	sub-agent picks each word’s gloss sense	−0.83
Self-revision	second pass edits the model’s own draft	−0.40
Ensemble selection	reference-free judge picks best of N drafts	+0.7 n.s.

Table 2: **Orchestration patterns tried beyond forced push (en→ht), all neutral-to-negative (Δ vs. the single-pass forced-push stack).**

3.3 Amplifier or equalizer? Grounding pays in inverse proportion to what the model already knows

Our results suggest that the harness acts as a reasoning *equalizer*, not an amplifier. The same evidence bundle that lifts a weak open model by up to +14 chrF++ into Creole leaves a strong frontier model flat: grounding does not scale ability up but substitutes for an ability the model lacks, and its benefit shrinks to essentially zero as the base model’s per-language competence grows.

The intuition first appeared comparing base against augmented qwen3-8b, qwen3-32b and gpt-4.1-mini; we broadened it to a wider panel, holding the harness fixed and varying only the LLM engine. On en→ht the same augmentation gives sharply different verdicts across a ten-model ladder over five families (Qwen, Mistral, Meta,

translator than critic, selector, or lexicographer.

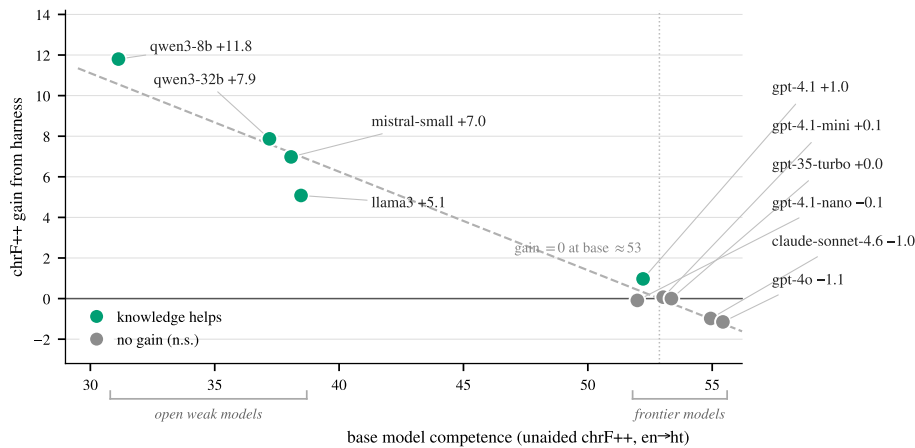


Figure 4: **Harness chrF++ gain vs. base-model competence (en→ht), ten models in five families (Δ = best-harness – base, first-200, matched idx; qwen3-8b: +11.8 first-200, +14 full-test).**

OpenAI, Anthropic) on a common first-200 slice: the four models below 40 chrF++ all gain heavily, while the six above 52 barely move (Figure 4). Each point is a 200-sentence estimate, so we read a robust *tendency* rather than an exact law, on a bimodal sample with no model between 40 and 52. Tools supply the Haitian-Creole vocabulary and surface patterns a weak model never acquired in training; on a model that already commands them, the same evidence is at best slight noise to reconcile against knowledge it trusts.

A study of the score distribution across several systems suggests that the harness lifts qwen3-8b’s whole en→ht distribution onto the primary baseline’s, chiefly by rescuing its worst sentences rather than shifting the curve uniformly, while the frontier’s advantage stays at the top (Appendix H).

Grounding an SLM with expert knowledge thus lets an 8B model starting at 28 chrF++ bridge the gap with a system fine-tuned for the language, but does not let it beat significantly larger models. To our knowledge no prior Creole-MT study has characterised knowledge injection as an equaliser whose payoff is predicted by the base model’s own literacy in the target language (shown here for Haitian Creole and the models we tested; other languages remain a hypothesis).

3.4 The cost of reasoning

Reasoning pays for a weak model, and only narrowly. Enabling native qwen3 chain-of-thought lifts ht→fr by +3.03 chrF++ (50.6 → 53.6),

curing the Creole-passthrough collapse of that French-lexified direction, but adds essentially nothing on the other three while costing roughly $1.7\times$ the tokens (its reasoning trace inflates output $\sim 24\times$ over a near-empty base, input unchanged; Figure 5). It is a targeted repair for one passthrough-prone direction, not a general gain.

Reasoning is gated by direction, and Table 3 must be read accordingly. The shipping configuration enables reasoning for ht→fr only. Into Creole it costs rather than pays, because reasoning and retrieval turn out to be substitutive rather than additive and retrieval reaches a higher score at a fraction of the cost, so we buy that gain with retrieval and leave reasoning off. Table 3’s positive into-Creole deltas are therefore not the effect of enabling reasoning: they measure the other levers within the reasoning-on arm, against a reasoning-on baseline.

That exception aside, the gains cluster where compute is cheapest. Knowledge-injection levers that add evidence to a single forward pass, at roughly $1\times$ base inference cost, carry almost all of the improvement. Orchestration and test-time-compute levers are flat or negative at every price: the agentic loop spends about $6\times$ the tokens for no gain, selecting among four drafts with a pairwise judge also costs $6\times$ and is not significant, two-pass quality-estimation selection costs $1.9\times$, and self-revision is cheap at $1.3\times$ and still loses ground (§3.2). Table 3 gives the per-lever verdicts, Appendix I the full per-tool

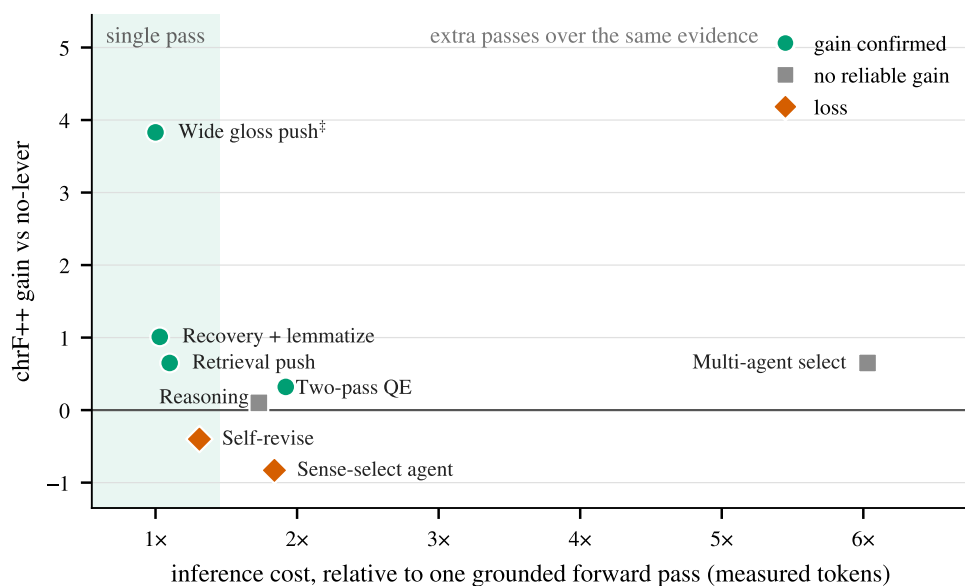


Figure 5: **Isolated single-lever chrF++ gain vs. measured inference cost (en→ht, first-200).** Each point is one lever’s own A/B; ‡ is plotted at single-pass marginal cost (isolated A/B $1.66\times$).

Lever	Direction	base $\rightarrow$ lever	Δ	Status
<i>Accepted (in shipping config)</i>				
retrieval_push ($k=3$)	en $\rightarrow$ ht	41.52 $\rightarrow$ 42.65	+1.14	kept
	fr $\rightarrow$ ht	41.27 $\rightarrow$ 42.70	+1.43	kept
enable_reasoning	ht $\rightarrow$ fr	50.57 $\rightarrow$ 53.60	+3.03	kept
gloss_bridge_lemmatize	en $\rightarrow$ ht / fr $\rightarrow$ ht	—	+1.30 / +0.54	kept
gloss_pivot (fr)	fr $\rightarrow$ ht	—	+4.56 [‡]	kept
gloss_skip_common=F	en $\rightarrow$ ht	—	+3.83	kept
grammar_push	fr $\rightarrow$ ht	—	+0.77 [‡]	kept
<i>Measured but superseded (not in shipping config)</i>				
recovery + lemmatizer ^a	en $\rightarrow$ ht	40.46 $\rightarrow$ 41.94	+1.48	superseded
(+ retrieval_push for en $\rightarrow$ ht)	fr $\rightarrow$ ht	40.73 $\rightarrow$ 41.43	+0.69	superseded
<i>Rejected (negative or non-significant; not shipped)</i>				
grammar_push	en $\rightarrow$ ht	—	−0.30	redundant
sense_select	en $\rightarrow$ ht	—	−0.83	hurts
revise (2nd pass)	en $\rightarrow$ ht	—	−0.40	hurts
gloss_collocations	en $\rightarrow$ ht	—	−0.06	n.s.
gloss_bridge_variants	ht $\rightarrow$ en / ht $\rightarrow$ fr	—	+0.02 / −0.05	n.s.

Table 3: **Per-lever ablation (Δ chrF++, signed)** against the baseline in each row’s *base $\rightarrow$ lever* column. $n=1010$; [‡] first-200; ^a inside the reasoning-on arm, not the effect of enabling it (§3.4). Detail: App. I.

breakdown, and Figure 5 makes the clustering visible. Injected knowledge at $1\times$ pays; buying more inference-time compute or orchestration does not.

3.5 Bottleneck is not data coverage but its correct exploitation

Studying the system trace helps us better understand how the LLM processes the injected knowledge. The model reads the dictionary reliably but exploits it unevenly; the limit is uptake, not resource breadth. Two observations support this.

First, gloss coverage is stable across the domain shift from FLORES+ to the blind challenge set. On three of four directions the covered-token rate barely moves: en $\rightarrow$ ht 89.0 $\rightarrow$ 86.6 (−2.4), ht $\rightarrow$ en 84.3 $\rightarrow$ 82.9. Only ht $\rightarrow$ fr drops sharply (−7.0), and there the uncovered words concentrate in the challenge set’s web-scrape slice: product listings, place and brand names, alphanumeric part-codes and foreign entities. The dictionary is thus not the bottleneck out-of-domain either.

Second, injected content is occasionally counter-productive. On the en $\rightarrow$ ht submission run against the same base with no harness, the harness lifts the mean by +12.5 chrF++, yet 7.7% of sentences score below the raw base, seven by more than ten points. The error typology shows the mechanism (Appendix J): the model grabs a wrong sense buried in a noisy Freeman entry (“neglected” $\rightarrow$ “madoulè”, “per cent” $\rightarrow$ “santim”), or the extra gloss and few-shot context perturb a sentence the base already handled.

This suggests an 8B base is simply not powerful enough to correctly exploit the injected knowledge.

Limitations

Five limitations bound the claims of this paper:

First, and most consequentially, our headline comparison is not size-matched. We compare a fine-tuned baseline of roughly two billion effective parameters against a harness-augmented eight-billion-parameter model, and we ran no matched-size control. We could not obtain the baseline’s Gemma base to use as our own backbone, and we could not fine-tune our 8B model under the compute and time available. Our claim that a training-free harness “beats trained” is therefore bounded by that mismatch: part of the gap we attribute to knowledge injection could, in principle, be a parameter-count effect. We flag this openly as the study’s central caveat.

Second, our metric is character-only. We report chrF++ throughout because it is what the shared task judges, but a surface-character metric penalises a correct synonym or an acceptable paraphrase as if it were an error. An obvious remedy, a trained neural quality estimator such as CometKiwi (Rei et al., 2022), is unavailable in practice: a reference-free COMET-QE model does not transfer to Haitian as it rewards source-adequacy and so scores a fluent copy of the source above a correct Haitian translation. We are thus limited to measuring quality through a character lens.

Third, the resource gap is structural, not incidental. Much of the Creole world is spoken far more than it is written, and the digitised lexica, grammars, and parallel corpora a knowledge-injection harness feeds on are thin for Haitian. Our results

are conditioned on the resources that happen to exist as we witnessed attempting to reproduce this approach with other Creole languages. Sometimes the resource exists but is not available, for lack of a digital version or for copyright reasons. Obtaining the linguistic resources for low-resource languages is the bottleneck, same as for the parallel corpora for training.

Fourth, the action space to explore with agentic approaches and context engineering is effectively infinite. One can always add another resource, re-format the injected evidence, or reframe the prompt, and any such move might shift a score. What we report are therefore systematic pairwise trends, each isolated under the protocol of §2.3, which exists because levers that looked like clean wins on 200 sentences repeatedly failed to confirm on the full set. We do not claim the magnitudes are portable to every model, resource base, or benchmark.

Fifth, the equalizer curve (Section 3.3) holds the harness fixed at the configuration we tuned for qwen3-8b and swaps only the model beneath it. Its verdict is therefore config-conditional: what we show is that qwen3-8b’s optimal harness does not help a frontier model, not that no harness would and a frontier-specific configuration might recover a small gain we did not look for. This result, on top of being model-specific, could also be language-specific.

4 Conclusion and perspectives

The system we describe was assembled under a deliberate constraint: avoiding all training and enhancing a weak open model only with knowledge injected at inference time. Within that budget it overtakes the primary baseline on all four directions of our development set, though not on the blind test set. For a language like Haitian Creole, a frontier model already carries enough latent competence that the residual question is no longer only whether we can translate it well, but how cheaply we can close the remaining gap. The bottleneck, in other words, has shifted from quality to frugality, and a frugal, explainable pipeline of deterministic lexical and grammatical evidence is a competitive answer, because the injected content is inspectable rather than baked into opaque weights. However, this approach does not solve Creole MT: parametric approaches still lead the benchmark.

This complementarity of train-time and test-time approaches changes the role of the harness itself.

In our longer-run design it is not the deliverable but a trace-collection engine: every translation emits its reasoning, tool calls, retrieved evidence, output and reward, and those grounded traces are the raw material for model adaptation. The frugal system and the trained one are then two ends of one pipeline, the cheap, explainable teacher that produces supervision no weak model could generate on its own. Pei et al. (2025) offer an existence proof: forward-translating a 42,240-sentence monolingual Manchu corpus with an in-context system, then fine-tuning *mT5-small* ($\sim 300\text{M}$) on the synthetic pairs, lifts chrF from 8.38 to 30.62 — most of the way toward the 37.93 the best in-context system reaches, and that 37.93 is a frontier model prompted in context, not a fine-tuned one.

Finally, to the best of our knowledge the substitution we observe has not been characterised across language types. To probe this substitution law further, the same harness should be extended to many typologically diverse languages, to determine whether our findings (the negligible impact of tool-calling delivery, and the trade between a model’s base capability and the benefit of context engineering) hold across languages or are intrinsically language-dependent. On the evidence of this shared task, train-time compute remains prevalent for low-resource MT.

Ethics Statement

Linguistic resource availability. All linguistic resources we use are openly licensed: FreeMan (CC BY-NC-ND), FreeDict (GPLv2+/CC BY-SA), NLLB (ODC-BY) and Wikipedia/kaikki (CC BY-SA); we declined borrow-only and all-rights-reserved sources and used no shadow-library material.

Reproducibility. Our system’s translation outputs on the FLORES+ devtest set, for all four directions, are released at <https://github.com/Dinarque/chv-creole-mt-2026-flores-outputs>.

Data contamination. The task requires participants to report the share of test segments present in their training data. We train nothing, so the analogous exposure is what retrieval can surface. The shipping configuration serves NLLB-mined bitext only, which gives 0.7% exact source overlap with the challenge set and none with any reference target. The decisive exclusion is the prohibited

Kreyòl-MT *test* split, 67–87% identical to the challenge sources, which no channel ever loads. The five fixed few-shot pairs come from FLORES+*dev*, and every result we report is on FLORES+*devtest*, which has 0% overlap with the challenge set and with every corpus in the pipeline.

On the use of Generative AI. Generative AI, and more specifically an AI coding assistant, was used in this work, notably to accelerate data ingestion and formatting inside the harness and to help with the pagination of the LaTeX file and graphs.

References

- Joseph Bernard. 1980. La graphie du créole haïtien. *Études Créoles*, 3(1):101–105. Official text of the Haitian Creole orthography; the orthography was ratified by the communiqué of the Secrétairerie d’État à l’Éducation Nationale of 22 January 1980.
- Vicent Briva-Iglesias. 2025. [Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication](#). *Preprint*, arXiv:2504.12891.
- Vicent Briva-Iglesias and Gokhan Dogru. 2025. AI agents may be worth the hype but not the resources (yet): An initial exploration of machine translation quality and costs in three language pairs in the legal and news domains. *Preprint*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2402.03216. Bge-m3.
- Zheng Chen, Fandong Meng, Jie Zhang, Yingxue Zhang, and Jie Zhou. 2024b. CRAT: A multi-agent framework for causality-aware reasoning and translation with retrieval-augmented terminology. *arXiv preprint arXiv:2410.21067*.
- Michel DeGraff. 2007. Kreyòl Ayisyen, or Haitian Creole (Creole French). In John Holm and Peter L. Patrick, editors, *Comparative Creole Syntax: Parallel Outlines of 18 Creole Grammars*. Battlebridge Publications, London.
- Nicholas Faraclas, Arthur K. Spears, Elizabeth Barrows, and Mayra Cortes Piñeiro. 2010. Orthography. In Arthur K. Spears and Carole M. Berotte Joseph, editors, *The Haitian Creole Language: History, Structure, Use, and Education*, chapter 4. Lexington Books, Lanham, MD.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. TEaR: Improving LLM-based machine translation with systematic self-refinement. *arXiv preprint arXiv:2402.16379*.
- Bryant C. Freeman. 2010. *Haitian-English Dictionary*, 6th edition. Institute of Haitian Studies, University of Kansas, Lawrence, KS.
- Gemma Team, Google DeepMind. 2025. Gemma 4. Base model `google/gemma-4-E2B`; the WMT26 Creole primary baseline is the `pranavgupta/gemma4-creole-mt-2026-baseline-adapter` LoRA adapter on this base.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [DiPMT: Dictionary-based phrase-level prompting of large language models for machine translation](#). *Preprint*, arXiv:2302.07856.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*. ArXiv:2305.04118.
- Jonathan Hus and Antonios Anastasopoulos. 2024. [Back to school: Translation using grammar books](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Ahrii Kim. 2025. Multi-agentMT: Deploying AI agent in the WMT25 shared task. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 583–586. Association for Computational Linguistics.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395. Association for Computational Linguistics.
- Heather Lent et al. 2024. [CreoleVal: Multilingual multitask benchmarks for creoles](#). *Transactions of the Association for Computational Linguistics*, 12:950–978.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of EMNLP 2024*, pages 17889–17904. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.
- Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 31372–31422. Association for Computational Linguistics.

- Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*.
- Yongyu Mu, Abudurexiti Rehemani, Zhijin Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. [Augmenting large language model translators via translation memories](#). *Preprint*, arXiv:2305.17367.
- NLLB Team et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*. Introduces FLORES-200.
- Open Language Data Initiative. 2024. FLORES+. <https://github.com/openlanguageata/flores>.
- Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788. Association for Computational Linguistics.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation (WMT): Research Papers*, pages 186–191. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.
- 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 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. Association for Computational Linguistics.
- Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Oreen Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew D. Stutzman, Bismarck Bamfo Odoom, 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: Human Language Technologies (NAACL)*. Association for Computational Linguistics.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#). *Preprint*, arXiv:2008.00401. MBART-50.
- Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. [A benchmark for learning to translate a new language from one grammar book](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Jason Wei et al. 2022. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research (TMLR)*.
- Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. In *Proceedings of EMNLP 2024, System Demonstrations*, pages 131–141. Association for Computational Linguistics. ArXiv:2405.11804.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.

A Extended related work

A.1 Current state of Creole MT

As for many languages, the lack of quality training data hinders the development of efficient and reliable MT systems for Creole languages despite high demand and millions of speakers.

Creole languages are a very interesting field to explore our hypothesis for various reasons, the first being that they share many common features with their substrate languages, it is easier to analyze the model’s productions compared to typologically different languages with a distinct vocabulary basis. This makes auditing the systems and error analysis more achievable.

They are, for historical reasons, particularly well documented, as many scholars from the 19th century onwards developed rich linguistic resources such as grammars and lexicons. Official bilingualism also often makes parallel corpora available in institutional resources or on websites even if that would not be enough to train state-of-the-art models.

Despite this “privileged” situation among low-resource languages, Creole MT remains understudied. Even with formal similarities to the substrate language and sometimes mutual intelligibility among them, they face specific challenges such as culture-specific realities and a lack of written traces for some registers and domains.

Works on Creole MT followed the AI paradigm shifts with the development of encoder-decoder models built with the available parallel corpora in different Creoles to leverage their typological proximity.

- **CreoleVal** (Lent et al., 2024): a benchmark over 28 Creoles whose MT track releases CREOLEM2M, a many-to-many model for English and 26 Creoles.
- **Kreyòl-MT** (Robinson et al., 2024): the largest public Creole MT corpus (41 Creoles) with fine-tuned mBART and NLLB models; the strongest specialized baseline for Haitian and our main comparison.

These works, while reaching very satisfying scores on in-domain benchmarks, display some limitations:

- The quality of the translation depends on the size of the parallel corpus for each Creole, and the training data is heavily imbalanced.
- The quality of the training data is questionable. For Kreyòl-MT, the parallel data is heterogeneous. Some is extracted from the web, sometimes poorly aligned, containing abusive language; and many examples do not correspond to casual language (item descriptions, very short sentences, a preponderance of religious texts for which a correct translation by an LLM could come from memorization rather than genuine language integration).

New efforts such as the WMT shared task encourage further work on these languages, recently introducing the role of LLMs.

Our experiments indicate that these models are now outperformed by frontier LLMs on our dev set (Section 3), probably due to the presence of monolingual data and reference textbooks in the pretraining data, which, combined with transfer learning and native fluency, puts them ahead of specialized models. The bottleneck is therefore not only quality but also adaptation to specific settings (healthcare) and cost optimization.

We focus on Haitian Creole (*Kreyòl*), the most widely spoken Creole, with roughly 12 million speakers in Haiti and a large diaspora across the United States, Canada, and France; it has been co-official with French since the 1987 constitution. Typologically it is a French-lexifier, isolating language with essentially no inflectional morphology, SVO order, and preverbal tense–mood–aspect markers, which makes its output relatively easy to audit. It is comparatively well documented with a standardized orthography (Bernard, 1980; Faraclas et al., 2010), reference dictionaries and grammars, and very much alive online and in music, arts, and news, so monolingual and reference material is comparatively plentiful in LLM pre-training.

A.2 Exploring training-free strategies for low-resource MT

The lack of sufficient parallel training data for low-resource MT led to a clear specialisation in the MT field, with a series of works attempting to find strategies suited to low-resource languages.

Many of these approaches rely on data augmentation, using synthetic data to improve the training data and therefore the model's performance. These approaches show limited benefits as they are dependent on the quality of the synthetic data and rarely address the lack of lexical coverage. The use of multilingual models also improved model translation via transfer learning, but the benefits are limited if no languages close typologically were used and while it improves model fluency, it does not address the lexical gap either.

The extended use of LLMs for MT opened a new, training-free perspective. While language modeling was not initially aimed at machine translation, early LLMs developed the emergent ability (Wei et al., 2022) to perform translation for language pairs well represented in their training corpus. However, for languages and domains absent from the training data, quality remains poor. But LLMs can be used via instructions and the prompt can be enriched with contextual information or precise instructions. Many works explored the input flexibility of LLMs to improve MT in low-resource scenarios. These works (2022–2024) focused either on prompting strategies or, mostly, on knowledge injection.

The injected knowledge comes from different sources:

- Use of lexicon / glosses to bridge the lexical gap for low-resource languages or specialized terminologies (Ghazvininejad et al., 2023). The main challenge is to ensure sufficient coverage in the resource used to ensure matching of the terms present in the input, and there is an inherent difficulty in capturing idioms. The matching mechanism often involves other NLP tasks such as lemmatization and Word Sense Disambiguation. Some of this matching process is handled by external code but some rely on the LLM reasoning abilities. This however does not greatly influence fluency and collocational quality: bridging the lexical gap ensures that the right words are made available to the model, but not that they are combined into natural, idiomatic target sentences. The linguistic reference should also follow the language's natural evolution (a dictionary is the partial and fixed image of the state of the language at a given moment in time). It does not include idioms, colloquialism or regional variations. This is particularly relevant for Creoles where there are some differences between the official language (for Haitian, the orthography standardized by the IPN in 1979 Bernard, 1980) versus the Creole spoken or written casually. The Kreyòl of northern Haiti displays some differences from the one spoken in the capital. . . Traditional linguistic resources mostly fail to capture this diversity.
- **Translation Memory.** The idea is to use relevant parallel sentence examples not at training time but at inference time incorporating them in the model instruction (Mu et al., 2023). This approach boosts fluency and helps produce more idiomatic sentences and correctly use words in context but its efficiency depends on several parameters: the effective closeness between the new sentence to translate and the provided exemplars. Therefore it depends on the size of the aligned corpus and the efficiency is limited in out-of-domain scenarios (same issue as the training based approaches); the quality of the retrieval process, which can be lemma based (precise but risk of a lower recall) or semantic, inspired by the RAG architecture, with the risk of incorporating examples that are further from the input sentence, therefore an important calibration work of the search hyperparameters is mandatory to optimize its benefits on a specific corpus; the number of exemplars provided (hyperparameter) and the ability of the underlying model to actually benefit from these examples.
- **Morphological analyzers:** to help the model identify grammatical features from an inflected form. This is highly relevant for morphologically rich languages such as Sanskrit but less so for Creole, which shows no inflectional morphology and has only a few homonyms from different grammatical categories (e.g. *manje*, both 'to eat' (v.) and 'food' (n.)).

The benefits of these approaches depend naturally not only on the quality and breadth of the LLM used but also on the capabilities of the LLM itself, and more specifically on its natural literacy in the target (and to a lesser extent source) language and its reasoning and metalinguistic abilities. Indeed, if the base model was not trained on sufficient data from the target language it usually fails to produce grammatically correct utterances. Hence another series of work focused on the LLM linguistic abilities and their capacity to learn new languages (either underrepresented or fictitious) from grammar books. Incorporating grammar excerpts and rules in the context of the generative model can steer the production of well-formed utterances but the main challenge faced is identifying which rule(s) to inject in context

with regard to the source sentence. Most of these works were therefore more probing the linguistic and generalization abilities of LLMs rather than applying it directly to MT (Marmonier et al., 2025); see also (Tanzer et al., 2024; Hus and Anastasopoulos, 2024).

All of these techniques have been shown to individually boost a baseline model, but the interaction of these different forms of expert knowledge has rarely been investigated. In a recent paper, Pei et al. (2025) treat the prompt as a modular space and quantify the relative benefit of each type of linguistic knowledge, using Manchu as a test case. The authors leveraged four different resources or techniques (dictionaries, grammar books, parallel examples and CoT prompting) and carried out a methodical ablation study for Manchu. To the best of our knowledge, these approaches have not been attempted on Creole languages and it is already a contribution to measure their efficiency.

A.3 Agentic approaches to MT

Since 2024, LLMs’ inner workings have evolved to let them perform a broader range of tasks: first the shift from train-time to test-time compute, with reasoning models that integrate CoT capabilities and output intermediate reflection traces before outputting a final answer to their instruction. It seems particularly relevant for translation tasks where some arbitrages are necessary to pick the best translation for an expression or where the decision process can be decomposed in multiple steps. A next step was reached with the advent of tool-calling capabilities, where the model can autonomously call provided tools and reason about their output until the task is performed (ReAct; Yao et al., 2023). This tool-calling ability, in what is now called agentic AI, suggests the opportunity to combine the latest LLM reasoning abilities to orchestrate external knowledge injection and thus optimize translation quality.

The main shift is that it introduced a novel paradigm for MT, framing it as a “multi-step, modular and potentially collaborative process” (Briva-Iglesias, 2025). Hence our positioning of this work as a reflection on the shift from train-time to test-time compute adapted for AI translation. What has been explored is multi-agent MT, along several lines:

- *Agents as a translation agency*: role-specialized LLMs (translator, editor, proof-reader) organised like a human translation company, as in TransAgents for ultra-long literary text (Wu et al., 2024).
- *Knowledge- and error-driven pipelines*: a model first mines the knowledge or diagnoses the errors it needs, then translates or refines accordingly: e.g. the human-like strategy mining of MAPS (He et al., 2024), the estimate-then-refine loop of TEaR (Feng et al., 2024), and multi-agent debate (Liang et al., 2024); CRAT (Chen et al., 2024b) adds retrieval to resolve unknown terminology.
- *LLM as a corrector / reviser*: iterative self-refinement (Madaan et al., 2023) applied to MT, but mainly on well-documented languages where the base model was already fluent, with marginal gains, inconsistent across domains, languages and metrics.

The exploration of agentic approaches for MT is, however, rather disappointing. Much work framed as agentic is rarely agentic in the most recent sense (namely tool-calling and reasoning loops). In reality it is mostly a combination of several LLM calls with specialized prompts. It can to a certain extent be called a multi-agent system, but since each individual LLM has no agency and only performs the task assigned to it in a single call, “workflow” describes it better. Kim (2025) is the only work to honestly frame it as “prompt chaining”.

These approaches are also costly (Briva-Iglesias and Dogru, 2025): the aforementioned work reports that sequential and iterative agent workflows consume roughly five and fifteen times the tokens of a single-pass system, respectively, while topping no automatic-metric comparison and edging ahead in only a single human-evaluation category. What makes agentic AI specific, the ability to autonomously call tools and leverage their output, was not explored at all, and the picture for low-resource MT remains inconclusive. Are the results more promising for low-resource languages, provided that adequate tooling can be built with quality resources? Our goal is to use the challenge as an opportunity to explore this new paradigm and determine to what extent these new training-free approaches can substitute for or complement fine-tuning.

B Configuration options

The harness exposes every behaviour as an explicit per-run flag; the default configuration equals the last-known-good baseline (a change ships only after it wins an A/B). Table 4 is the reference key — what each lever means and its value in the eng→hat submission run; shaded cells are the deliberately active (non-baseline) settings that constitute the system.

Table 4: Meaningful harness configuration flags: meaning, dataclass default, and value in the eng→hat *challenge* submission config (`evaluation/submit_challenge.py`). Reasoning is direction-gated and off for eng→hat, so this is the reasoning-off arm (see `enable_reasoning`). Active non-baseline settings shaded green.

Flag	Meaning	Default	Shipped
Dictionary / gloss push			
<code>dictionary_variant</code>	Delivery mode for lexical evidence: <code>batch/single</code> expose a callable <code>dict_lookup</code> tool (ReAct loop); <code>off</code> = evidence force-pushed into a direct-translation prompt.	<code>batch</code>	<code>off</code>
<code>gloss_push</code>	Auto-gloss the source sentence and inject the glosses into the prompt (imposed, not self-served).	<code>False</code>	<code>True</code>
<code>gloss_skip_common</code>	Push rare words only; <code>False</code> = gloss common words too (weak models mistranslate common words as well).	<code>True</code>	<code>False</code>
<code>gloss_common_topn</code>	Frequency rank below which a source word counts as “common”.	<code>3000</code>	<code>3000</code>
<code>gloss_bridge_lemmatize</code>	On a look-up miss, resolve an inflected word to its headword with the right POS (called→call) before fuzzy recall.	<code>False</code>	<code>True</code>
<code>gloss_idioms</code>	Detect Freeman idioms/MWEs in the source and surface them (adds a line only on a hit).	<code>True</code>	<code>True</code>
<code>gloss_pivot</code>	For a source with no lexicon index (<code>fr</code>), pivot content words <code>FR→EN→hat</code> so the wide-gloss lever fires.	<code>False</code>	<code>False</code>
<code>gloss_clean</code>	Clean+rank raw gloss blobs into short best-first renderings (strip POS/sense/OCR noise).	<code>False</code>	<code>False</code>
Few-shot demonstrations			
<code>fewshot_k</code>	Fixed (source⇒target) example pairs prepended to the prompt, selected deterministically from the gold corpus (leakage-safe). <code>0</code> = off.	<code>0</code>	<code>5</code>
<code>fewshot_dynamic</code>	Retrieve the <i>k</i> demos most similar to the sentence instead of static even-spaced ones.	<code>False</code>	<code>False</code>
Dense retrieval push			
<code>retrieval_push</code>	Inject dense-retrieved parallel + monolingual-fluency exemplars (a dynamic domain few-shot). Into-hat only.	<code>False</code>	<code>True</code>
<code>retrieval_k</code>	Exemplars retrieved per channel.	<code>3</code>	<code>3</code>
<code>retrieval_min_sim</code>	Cosine floor (<code>bge-m3</code>) below which an exemplar is dropped.	<code>0.55</code>	<code>0.55</code>
<code>retrieval_mono</code>	Include the monolingual (fluency) retrieval channel alongside the parallel one.	<code>True</code>	<code>True</code>
<code>retrieval_drop_self_match</code>	Drop any exemplar whose source is verbatim the query sentence (blind-eval integrity guard).	<code>False</code>	<code>True</code>
Grammar			
<code>grammar_push</code>	Auto-inject top grammar rule-cards. Helps only where there is no gloss (<code>fr-hat</code>); redundant on <code>en-hat</code> .	<code>False</code>	<code>False</code>
<code>grammar_query_features</code>	Detect the sentence’s marked constructions (TMA, negation, questions) to steer card retrieval. Inert unless <code>grammar_push</code> is on, i.e. <code>fr→hat</code> only.	<code>True</code>	<code>True</code>
Prompt rules			
<code>no_codeswitch</code>	Translate every common word; do not leave/transliterate source words (keep proper names).	<code>False</code>	<code>True</code>
<code>keep_proper_names</code>	Copy names verbatim, never phonetically respell them.	<code>False</code>	<code>False</code>
<code>idiomatic_hint</code>	Render the meaning as a native speaker would; treat the glossary as reference, not a script.	<code>False</code>	<code>False</code>
Decoding / robustness			
<code>temperature</code>	Decoding temperature (greedy).	<code>0.0</code>	<code>0.0</code>
<code>enable_reasoning</code>	Emit chain-of-thought before the answer in one decode. Direction-gated: off into Creole, on for <code>ht→fr</code> (+3.03); when on, <code>max_tokens</code> → 8192 and <code>final_max_tokens_ratio</code> → 0, else the per-word cap truncates the CoT.	<code>False</code>	<code>False</code>
<code>final_max_tokens_ratio</code>	Anti-runaway cap on the final decode: tokens per source word. <code>0</code> = disabled.	<code>6.0</code>	<code>6.0</code>
<code>regenerate_on_collapse</code>	If the output is degenerate (repetition/length explosion), regenerate that sentence with a frequency penalty; keep the retry only if it scores better. Into-hat only.	<code>False</code>	<code>True</code>
<code>collapse_retry_drop_push</code>	On a collapse retry, rebuild the prompt without the retrieval block (cures retrieval-induced loops). Into-hat only.	<code>False</code>	<code>True</code>
<code>normalize_repetition</code>	Deterministic post-edit collapsing degenerate repeated n-grams; safety net, never touches fluent output.	<code>False</code>	<code>True</code>

Per-direction adaptation. Table 4 is the eng→hat configuration; the other three directions run the same stack and differ in four flags only, each following from evidence availability rather than from a per-pair search — French has no direct hat lexicon, the dense index and the collapse guard are Creole-side, and reasoning repairs the hat→fra passthrough collapse (§3.4).

Direction	Difference from the eng→hat configuration
eng→hat	— (the reference configuration, Table 4)
fra→hat	+ <code>gloss_pivot</code> , + <code>grammar_push</code>
hat→eng	− <code>retrieval_push</code> , − <code>regenerate_on_collapse</code> , − <code>collapse_retry_drop_push</code>
hat→fra	as hat→eng, + <code>enable_reasoning</code>

What was submitted. The blind-set files enable `enable_reasoning` in all four directions, so our official scores correspond to the + *harness* + *reasoning* row of Table 1, not the *shipped* row.

C Tool input/output

The three tools the ReAct agent can call: the description and argument schema as shown to the model (from `tools/*.py`, abridged), a *real* call and its real output (tool code run locally).

C.1 Lexicon — `dict_lookup`

`dict_lookup(word*, lang, reverse)` — individual word lookup

What the model sees. Look up one source word/lemma in the Freeman dictionaries (one word, not a phrase); forgiving on accent/spelling/clitic variants; returns candidate lemmas with senses (POS + gloss). **Args:** `word*` (str); `lang` (ht|en|fr); `reverse` (bool, gloss-side hits too).

```
>>> dict_lookup("shaft", lang="en")
Dictionary candidates for "shaft":
- shaft (n): n. 1 (mach.) chaf 2 (tunnel) tinèl, chemen anba td 3 (handle) manch ...
- shift (n/v) [edit]: n. 1 (change) chanjman, varyasyon 2 (work) (la)relèv, ekip ...
>>> dict_lookup("barb", lang="en")
Dictionary candidates for "barb":
- barb (n): n. 1 (thorn) pikan 2 (insult) pawòl piman bouk, afwon
- barbed (adj) [edit]: adj. bable; ~ wire filfè, fildefè bable [Fr.]
```

C.2 Translation memory — `tm_search`

`tm_search(sentence*, k)` — fuzzy bitext retrieval

What the model sees. Fuzzy-match the source against the bitext and return the closest past translations as `source => target` with a chrF++ match %. **Args:** `sentence*` (str); `k` (int, default 3).

```
>>> tm_search("Wash your hands with soap and water.", k=3)
Similar past translations (source => target, % = chrF++ match):
- [87.2%] Wash hands with soap and water. => Lave men ou avèk dlo ak savon.
- [48.7%] Wash your hands frequently with soap and water for at least 20 seconds.
    => Lave men ou souvan avèk savon ak dlo pandan 20 segonn.
```

In-domain (health) input hits; on encyclopedic FLORES sentences the same call usually returns nothing.

C.3 Grammar — `grammar_lookup`

`grammar_lookup(sentence*, k)` — rule-card retrieval

What the model sees. Retrieve Haitian Creole grammar rule cards for a sentence — preverbal TMA markers (`te/pral/ap`), postposed determiners, negation `pa`, `se` clefts, possession. **Args:** `sentence*` (str); `k` (int, default 3).

```
>>> grammar_lookup(source, k=2)
Grammar notes (morphosyntax):
- [TMA: past / anterior (te)] Preverbal 'te' marks anterior; subject + te + verb; no verb
  inflection for tense. e.g. Li te manje. -- He ate / had eaten.
- [Negation: 'pa' placement] 'pa' precedes any TMA marker + verb: pa te, pa ap (pap),
  pa pral. e.g. Mwen pa konnen. -- I don't know.
```

D Linguistic resources used

Inventory of the Haitian Creole (ht) resources the harness draws on, grouped by **tool destination** (which channel consumes them) and listed alphabetically within each group. The colour code marks whether a resource is loaded by the **final submission config** (final), was queried in an ablation only (exp.), or was considered but **not retained** (decl.), with the reason (copyright or data quality) given in-line. The paper is ht-first; the four study languages sit on the Wiktionary/kaikki floor, as discussed in the main body.

final in submission config exp. ablation only decl. considered, not retained

Resource	Size / role	License	Source	Used
<i>Gloss lexicon</i> (dictionary tool, both directions)				
Freeman & Laguerre, <i>HT-EN Dict.</i> , 6th ed. (2010)	78,843 entries (index 80,704, +16k idioms); reference gloss lexicon	CC BY-NC-ND 4.0	KU ScholarWorks 1808/11816	final
Haiti-Référence online dict.	13k entries (12k FR glosses); direct HT→FR oracle — <i>copyright</i> , dev-aid only	all rights reserved	haiti-reference.info	decl.
Freeman & Laguerre, <i>HT-EN Medical Dict.</i> (1997)	1,110 entries; medical domain	CC BY-NC-ND 4.0	KU ScholarWorks 1808/10892	final
Védrine, <i>leksik kreyòl</i>	25 entries; terminology glossary	free (author-hosted)	Potomitan (Védrine)	final
Heurtelou & Vilsaint, <i>Diksyonè Medikal</i> (2000)	2,978 entries; medical domain — <i>copyright</i> , cannot use	all rights reserved	Educa Vision (in print)	decl.
Targète & Urciolo, <i>HT-EN Dict.</i> (1993)	7,611 entries; secondary bilingual seed — <i>copyright</i> , cannot use	all rights reserved	Dunwoody Press (in print)	decl.
Valdman et al., <i>HCEBD</i> (2007)	≈30k entries; gold-standard — <i>copyright</i> , borrow-only (CDL)	all rights reserved	IU Creole Institute	decl.
Vilsaint, <i>Diksyonè kreyòl-anglè</i>	34k terms; commercial — <i>copyright</i> , preview-only	all rights reserved	Educa Vision	decl.
<i>Pivot lexica</i> (fr ↔ ht gloss, via English)				
FreeDict <i>eng-fra</i> v0.1.6	8,799 headwords; EN→FR leg of ht→fr gloss	GPLv2+ / CC BY-SA	FreeDict	final
FreeDict <i>fra-eng</i> v0.4.1	8,505 headwords; FR→EN leg of fr→ht pivot	GPLv2+ / CC BY-SA	FreeDict	final
kaikki French (Wiktionary)	368,109 EN-headed entries; enlarges fr→ht pivot	CC BY-SA 3.0	kaikki.org	final
<i>Parallel corpora</i> (dense retrieval / TM)				
FineTranslations	web-scale; <i>synthetic</i> (Gemma3 MT target) — <i>data quality</i> , not gold	ODC-BY	HuggingFaceFW	decl.
ISI many-eng v2	~52k novel pairs; <i>data quality</i> — CCAIghed web-junk	per-source	rtg.isi.edu/many-eng	decl.
Kreyòl-MT (Robinson et al., 2024)	hat-eng 665k, hat-fra 353k; TM / mBART-baseline probe only	per-source aggregate	Robinson et al. (2024)	exp.
NLLB-mined	ht-en / ht-fr 300k/300k; 30k dense retrieval index (into-ht)	ODC-BY	allenai/nllb (HF)	final
VOA Creole news	~23k PD articles; <i>data quality</i> — no article-level pairing	mixed (PD + wire)	voanouvel.com	decl.
<i>Monolingual</i> (target-side fluency channel)				
Wikipedia Haitian Creole (2023-11-01)	30k-sentence sample; bge-m3 fluency channel (into-ht)	CC BY-SA	Wikimedia dump	final
<i>Grammar and orthography</i> (grammar-RAG / normalization)				
DeGraff (2007), <i>Kreyòl Ayisyen</i>	descriptive grammar; source of grammar-RAG cards	self-archive; un-verif.	MIT (author)	final
IPN orthography (1979)	official phonemic spelling standard; normalization target	standard (n/a)	Institut Pédagogique National	final

Table 5: Haitian Creole resources, grouped by tool destination. Idiom sub-entries fold in via `build --idioms`; the shipping gloss index totals 70,859 entries after OCR / de-columnization repair.

E Results comparison across systems

Three FLORES+ devtest sentences (en → ht) through five systems: **baseline** (Gemma-4-E2B + Creole-MT LoRA), **qwen** (no harness), **+harness** (forced push, submission config), **+harn+reason**, **frontier** (gpt-4.1-mini); the **src** and **gold** rows sit in the output column, above the outputs. Output words are matched to the gold *in order* (longest-common-subsequence, tolerating a displacement of up to three words): matched words are **green**, and a word that differs from its aligned gold word only in accents or case, or by a single character, is **yellow**; punctuation is ignored. chrF++ is character-level and *case-sensitive*, so yellow words still cost score. chrF++: **≥55** **40–54** **<40**.

system	chrF++	output / reference
idx 30		
src		Because the dinosaur feathers do not have a well-developed shaft, called a rachis, but do have other features of feathers — barbs and barbules — the researchers inferred the rachis was likely a later evolutionary development that these other features.
gold		Kòm plim dinòzò yo pa gen yon tij byen devlope, ki rele rachi, men yo gen lòt karakteristik plim — bab ak babil — chèchè yo dedwi ke rachi a te pwobableman yon devlopman evolutif ki vini apre lòt karakteristik yo.
baseline	36.9	Pwenn ren asay dinoza yo pa gen yon ti tisi long tankou ras la, men li gen lòt karakteristik nan pwenn tankou rad la ak barbule yo, pou sa rekeyachè yo te konklize ras la te gen tandans yo t ap devlope pita atravè lòt karakteristik sa yo.
qwen	35.2	Kèk kominote dinazaur yo pa gen yon shaft bien developman, ki apelé yon rachis, men yo gen yon fèt lòt ki sèlman kominote — barb ak barbül — rechèch yo konklize ke rachis yo te pwosedi yon developman evolwayonè lèt ki sa yo.
+harness	53.0	Paske zwazo dinòzò pa gen yon chaf bien devlope, ki apelé yon rachis, men yo gen lòt karakteristik zwazo — pikan ak barbül — chichi yo dedwi ke rachis te pwobabman yon evolisyonnis devlope pita k sa lòt karakteristik yo.
+harn+reason	42.0	Paske pm il dinòzò yo pa gen yon chaf bien devlope, ki te rele rachis, men genyen lòt kalite pm il — pikan ak pikan minil — chichi yo te dedwi ke rachis te genlè yon evolisyonnis developman ki te pita pase sa lòt kalite yo.
frontier	72.1	Paske plim dinòzò yo pa gen yon baton byen devlope, yo rele l yon rachis, men yo gen lòt karakteristik plim — barbs ak barbules — chèchè yo dedwi ke rachis la te pwobableman yon developman evolisyonè ki te vini apre lòt karakteristik sa yo.
idx 39		
src		During his trip, Iwasaki ran into trouble on many occasions.
gold		Pandan vwayaj li, Iwasaki te gen plizyè pwoblèm.
baseline	60.4	Pandan vwayaj li a, Iwasaki te rankontre pwoblèm anpil tan.
qwen	30.9	Pou li te fè yon pèw, Iwasaki te ranpli kounye a nan peryò diferan.
+harness	61.5	Pandan vwayaj li, Iwasaki te eksplike anpil okazyon.
+harn+reason	67.2	Pandan vwayaj li, Iwasaki te kounye a pwoblèm anpil okazyon.
frontier	75.0	Pandan vwayaj li, Iwasaki te jwenn pwoblèm plizyè fwa.
idx 185		
src		France is the seventh country in the European Union to suffer this virus; following Austria, Germany, Slovenia, Bulgaria, Greece and Italy.
gold		Lafrans se setyèm peyi nan Inyon ewopeyèn k ap soufri ak viris sa, apre Otrich, Almay, Esloveni, Bilgari, Grès ak Itali.
baseline	47.4	Frans se setèm peyi nan Inyon Ewopeyen an pou soufize a maladi sa a; sa a te swiv Otrich, Almay, Sloveni, Bulgarie, Gris ak Itali.
qwen	23.9	Frans la se septèm pwovans yo nan Unyon Europeen ki te fèt vire yo; apre Austri, Alman, Slovene, Bulgar, Gres e Italyen.
+harness	76.2	Frans se setyèm peyi nan Inyon Ewòpeyen ki soufri viris sa; apre Otrich, Almay, Esloveni, Bilgari, Grès ak Itali.
+harn+reason	67.2	Lafrans se setyèm peyi nan Inyon Ewdpeyen ki soufri sa viris sa; apre Otrich, Lalmay, Esloveni, Bilgari, Lagrès ak Itali.
frontier	63.6	Lafrans se setyèm peyi nan Inyon Ewopeyen an ki soufri viris sa a; apre Otrich, Almay, Sloveni, Bulgari, Lagrès ak Itali.

Proper names are not exempt. The released FLORES+ Haitian reference for idx 39 reads `lwasaki` where both source columns read `Iwasaki`, an I/I confusion in the gold itself; the scores above use our patched local copy.

F Official shared-task results

Official scores on the task’s blind test set for our four directions. Our three submitted systems were the bare model (`contrastive2`), the harness (`contrastive1`) and the harness with reasoning (`primary`). Δ is against the primary baseline (Gemma-4-E2B + Creole-MT LoRA); rank is among all entries scored in that direction, baselines included. **Bold** = best in the direction; **green** = our best submission.

System	chrF++	Δ	Rank
en→ht			
Kreyòl-MT (secondary baseline)	37.44	+4.41	1/6
best other participant (<code>gmu</code>)	34.62	+1.59	2/6
Gemma + LoRA (primary baseline)	33.03	—	3/6
ours: qwen3-8b + harness + reasoning	30.16	−2.87	4/6
ours: qwen3-8b + harness	30.10	−2.93	5/6
ours: qwen3-8b, no augmentation	21.87	−11.16	6/6
fr→ht no other team entered			
Gemma + LoRA (primary baseline)	31.31	—	1/5
Kreyòl-MT (secondary baseline)	31.11	−0.20	2/5
ours: qwen3-8b + harness	24.79	−6.52	3/5
ours: qwen3-8b + harness + reasoning	24.68	−6.63	4/5
ours: qwen3-8b, no augmentation	19.42	−11.89	5/5
ht→en			
best other participant (<code>pik</code>)	50.70	+11.63	1/10
Kreyòl-MT (secondary baseline)	40.97	+1.90	5/10
Gemma + LoRA (primary baseline)	39.07	—	7/10
ours: qwen3-8b + harness	35.65	−3.42	8/10
ours: qwen3-8b + harness + reasoning	35.57	−3.50	9/10
ours: qwen3-8b, no augmentation	33.95	−5.12	10/10
ht→fr			
best other participant (<code>uqam</code>)	42.43	+8.48	1/8
Gemma + LoRA (primary baseline)	33.95	—	3/8
Kreyòl-MT (secondary baseline)	33.70	−0.25	4/8
ours: qwen3-8b + harness + reasoning	31.06	−2.89	5/8
ours: qwen3-8b + harness	28.69	−5.26	6/8
ours: qwen3-8b, no augmentation	28.23	−5.72	7/8

Three readings. We finish below both baselines in every direction, and behind every other team that entered the same one. The harness itself still works on unseen data: it adds +8.23 (en→ht) and +5.37 (fr→ht) over the bare model against +1.70 and +0.46 out of Creole, and reasoning is worth +2.37 in ht→fr and nothing elsewhere, both as on our dev set. Every system scores far lower here than on FLORES+ devtest (ours −11.7 to −22.5 chrF++, Gemma −7.4 to −17.1, Kreyòl-MT −9.9 to −18.6), so our dev-set margin over the primary baseline does not survive the domain change.

G Anatomy of a run

One en → ht sentence (FLORES+ devtest, idx 30) shown as a **complete run in both delivery modes**; verbose evidence blocks are abridged for space. The base model (qwen3-8b) and the tools are identical across modes; only the delivery differs. The forced-push prompt is a nest of blocks, one shell holding a **SYSTEM** and a **USER** prompt, each assembled in a fixed order.

base template filled variable few-shot lexical gloss retrieval

G.1 Forced push

Forced-push prompt

single LLM call • blocks in shipping-assembly order

1 • SYSTEM

per-direction • prompt-cache-friendly

a base template

You are an expert **English**→**Haitian Creole** translator **specialised in Haitian Creole**. Translate the user's **English** sentence into accurate, natural **Haitian Creole**, using **Haitian Creole's IPN-1979 orthography**. Output **ONLY** the translation -- every ordinary word in **Haitian Creole**; never leave or transliterate a foreign word (e.g. "speeds" -> "vitès").

b static few-shot exemplars (5 total; 2 shown)

English: In all cases, you must book by phone directly with the airline.
Haitian Creole: Nan tout ka yo, ou dwe rezève nan telefòn dirèkteman bò kote konpayi ...
English: Dump trucks were used to block tube entrances ...
Haitian Creole: Kamyon baskil yo te itilize pou bloke antrè tib ...

2 • USER

per-sentence • assembled top-to-bottom in order

a task + source sentence

Translate this **English** sentence into **Haitian Creole**:
Because the dinosaur feathers do not have a well-developed shaft, called a rachis, but do have other features of feathers -- barbs and barbules -- the researchers inferred ...

b lexical gloss 18 glosses + 5 idioms (candidate senses, verbatim; 10 shown)

Optional lexical hints for less-common words (candidate senses -- use only if helpful; prefer natural, idiomatic phrasing and do NOT translate word-for-word):
Glossary (source word -> candidates):
- because (conj): conj. paske, pase, diske, piske, akòz (de), dapre, poutet ~ of ...
- feather (n): n. pm il (zwazo) ~ neither God nor man san fwa in fine ~ frekou kola ...
- do (v): v. fe - it! baye!, koupe kloure! don t ~ that! fok ou pa fe sa! ...
- have (m/v): v. genyen, gen I don t ~ anything M pa genyen anyen to ~ (beard) ...
- develop (v): v. 1 (to evolve) devlope, evolve, tounen, dewoule 2 (fix up) amenaje ...
- shaft (n): n. 1 (mach.) chaf 2 (tunnel) tinèl, chemen anba tè 3 (handle) manch ...
- but (conj): conj. 1 (however; yet; still) men, sepandan, tanselman 2 (only) selman ...
- feature (n/v): n. kalite, karakteristik [Fr], kalbas, kannannan, mak fabrik ...
- barb (n): n. 1 (thorn) pikan 2 (insult) pawol piman bouk, afwon
- later (adv/n): adv. pita, yon lot ... ~ on pi devan; see you ~ n a we pita
... + 8 more (18 glosses total)
Idioms / multiword expressions (prefer these over word-by-word):
- pa pile si li te konnen -- to be careful so as not to have later regrets
- woule nan/sou lwil -- to be rolling in clover, have it made, be well-off
... + 3 more (5 idioms total)

c retrieval few-shot -- into-Creole, k=3 dynamic domain exemplars

Similar sentences (source => reference target), nearest first:
<k=3 nearest bitext pairs, injected for into-Creole directions -- abridged>

MODEL — Final Answer (1 call)

chrF++ 55.7

Paske plim dinozò yo pa gen yon chaf bien devlope, ki apelé yon rachis, men yo gen lòt karakteristik plim -- pikan ak pikan minil -- chichi yo te infere ke rachis yo genlè pwobab li te yon evolwiyonnis devlope pita ki sa lòt karakteristik yo.

The output adopts the pushed glosses almost verbatim: **dinozò, plim, chaf, karakteristik, pikan, evolisyonnis, chichi, genlè, pita**.

G.2 Agentic (ReAct)

The same instruction, but the tools are now *callable* and native qwen3 reasoning is on: the model runs a Thought → Action → Observation loop until it answers (two LLM calls).

Agentic prompt — SYSTEM

base template + callable tools

a base template

You are an expert English->Haitian Creole translator specialised in Haitian Creole. Your job: translate the user's English sentence into accurate, natural Haitian Creole, using Haitian Creole's IPN-1979 orthography.

b callable tool interface (abridged)

You may consult tools before answering:
- dict_lookup(word, lang): one source word -> candidate lemmas with senses (POS + gloss).
- tm_search(sentence, k): closest past translations, 'source => target' with a match %.
- grammar_lookup(sentence, k): Haitian Creole grammar rules (TMA markers, determiners, ...).
Format -- Thought: <reasoning> Action: <tool> Action Input: <JSON> ; when ready, Thought: <final> Final Answer: <the Haitian Creole translation only>.

THOUGHT 1 (plan)

First I'll identify the technical terms -- "rachis", "barbs", "barbules" -- and check whether Haitian Creole has established translations or keeps them as is.

→ ACTION dict_lookup

Action Input: {"word": "rachis", "lang": "ht"}

← OBSERVATION 1 candidate (edit-distance match)

Dictionary candidates for "rachis":
- rach (n) [edit]: n. ax, hatchet; Ax (for felling trees). Cf. Hatchet.

REASONING native qwen3 chain-of-thought (condensed)

The lookup returned "rach" (an ax) -- unrelated; a rachis is a feather shaft. I'll keep the technical terms as is and translate the rest.
[It then fabricates "barbs"/"barbules" lookups it never issues, and drafts from its own knowledge.]

MODEL — Final Answer 2 calls, reasoning on

chrF++ 31.7

Kèk kounye a, poulè dinosaù yo pa gen yon shaft bien dévloppe, ki apelé rachis, ma yo gen lòt fèt yo--barb ak barbule--èske konesans yo te pran kounye a ke rachis te yon developman evolwayon ki te pase sa lòt fèt yo.

The agent’s one lookup fuzzy-matches rachis → rach (“ax”) and rightly discards it, then leaves shaft, barb and barbule untranslated and hallucinates the rest; reasoning recovers none of it. Forced push (above) had handed it plim, chaf, pikan and dinozò up front and scored chrF++ 55.7 against 31.7. Same model, same tools; only the delivery changed.

G.3 Reference and overlap

Gold reference — FLORES+ ht (idx 30)

Kòm plim dinozò yo pa gen yon tij byen devlope, ki rele rachi, men yo gen lòt karakteristik plim -- bab ak babil -- chèchè yo dedwi ke rachi a te pwobableman yon devlopman evolutif ki vini apre lòt karakteristik yo.

Forced push adopts the top *pushed gloss* — which only sometimes equals the gold word: =gold exact match, syn a valid Creole synonym, OCR? a likely Freeman OCR artifact; the agent instead leaves English / wrong sense .

source	gold ref.	forced push (= pushed gloss)	agentic
dinosaur	dinozò	dinozò	dinosaù
feather	plim	plim	poulè
shaft	tij	chaf	shaft
barb	bab	pikan	barb
researcher	chèchè	chichi	konesans
later	apre	pita	—

The win is not that the gloss equals the gold, only dinozò and plim do. It is that forced push reliably yields *Creole* (a gold word, a synonym, or even a noisy gloss), which chrF++ rewards, whereas the agent leaves English (shaft, barb). chichi for “researcher” is copied faithfully even though it looks like an OCR corruption of chèchè.

H Score distribution across systems

Does the harness lift qwen3-8b’s whole per-sentence score distribution onto the primary baseline’s, or only its mean, and is the floor (the worst sentences) equalized or still poor? Figure 6 compares five systems on one en→ht FLORES+ devtest index set ($n=1010$).

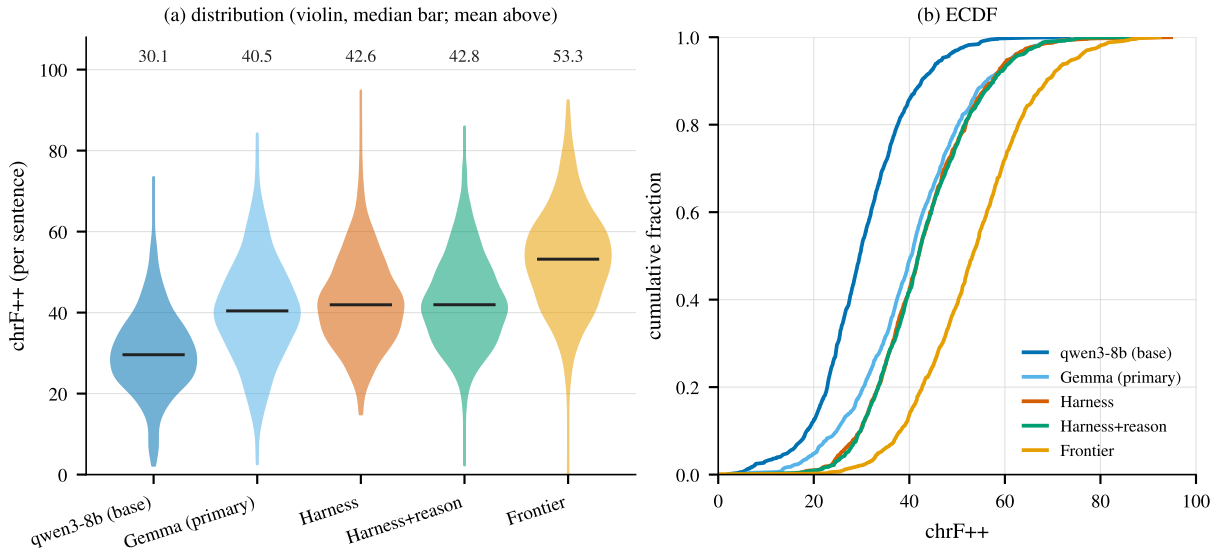


Figure 6: Per-sentence chrF++ distributions across five systems, en→ht (common index, $n=1010$): (a) violins, median bar and mean above; (b) ECDFs.

The harness shifts the bare SLM’s distribution right onto the Gemma baseline’s (means 42.6 vs. 40.5), with mild compression rather than a reshape (standard deviation of 10.9, between base 9.8 and Gemma 12.3). The gain concentrates at the floor: sentences below chrF++ 20 fall from 12.3% (base) to 0.9% (under Gemma’s 4.8%) and the base’s worst-decile mean rises 13.0 → 36.3 (Gemma 35.1). So the equalizer raises the worst sentences to baseline level rather than translating the curve uniformly; reasoning adds nothing, and the residual gap to the frontier is entirely at the top (median 53).

I Ablation by tool

This appendix works through the harness one lever at a time. For each variant we give the chrF++ delta against the stated baseline (green = a confirmed gain). Each row also names the direction it was measured in (**Dir.**) and the eval set it was measured on (**Set**): `full-1010` is the whole devtest, while `first-200` and `isolated` are the smaller slices we use for cheaper estimates. Because the two kinds of run are mixed here, compare deltas within a lever rather than across sets. All numbers are chrF++ (sacrebleu). These are the runs behind Table 3 in the body.

I.1 Dictionary / gloss push

The core lever: content-word glosses (Freeman HT↔EN, plus mined idioms) injected into the prompt.

Variant	Dir.	Set	Δ chrF++
<code>gloss_skip_common=False</code> (wide)	en→ht	isolated	+3.83
<code>gloss_idioms=True</code> (default on)	en→ht	–	(kept; no isolated A/B)
<code>gloss_collocations=True</code>	en→ht	full-1010	–0.06 (n.s.)
<code>sense_select</code> (vocab sub-agent)	en→ht	full-1010	–0.83

Using glosses: narrow 35.58 → wide 39.42, isolated. Collocations are too rare on FLORES; the same-model vocab sub-agent hurts (re-thinking underperforms injecting). Vocabulary delivery mode matters too: agentically, a batch lookup gives qwen3-8b +4.29 (sig., $p=.004$) vs. +0.83 (n.s.) for one-word-per-turn, at $\sim 6\times$ lower cost (en→ht, $n=60$).

I.2 Source lemmatizer / gloss bridge

A POS-aware source lemmatiser (NLTK + WordNet) resolves inflected source words to the headword with the correct sense (*called*→call, *found*→find→jwenn); the from-ht analogue adds coverage but no score.

Variant	Dir.	Set	Δ chrF++
<code>gloss_bridge_lemmatize</code>	en→ht	full-1010	+1.30 ($p<.001$)
<code>gloss_bridge_lemmatize</code>	fr→ht	full-1010	+0.54 ($p<.001$)
<code>gloss_bridge_variants</code>	ht→en	–	+0.02 (n.s.)
<code>gloss_bridge_variants</code>	ht→fr	–	–0.05 (n.s.)

Into-ht deltas are combined with the +2,762-entry Freeman lexicon recovery. From-ht is neutral (the target en/fr is high-resource, not source-coverage-limited).

I.3 FR→EN→HT pivot gloss

French has no on-disk lexicon, so we pivot fra→eng (FreeDict) then eng→hat (Freeman) to give fr→ht the wide-gloss lever it lacked. `gloss_pivot` (FR→EN→HT), fr→ht, first-200: +4.56 (34.52 → 39.08, CI [3.20, 5.77], $p\approx 0$) — the largest single fr→ht lever.

I.4 Dense retrieval push (into-ht)

For each source sentence we retrieve the three most similar sentence pairs from the NLLB-web parallel data (bge-m3 cosine, floor 0.55), plus a few Creole-only sentences, and put them in the prompt as worked examples. What the model gets is a fresh set of examples of how the target language is actually written, chosen per sentence instead of fixed in advance.

Variant	Dir.	Set	Δ chrF++
<code>retrieval_push k=1</code>	en→ht	full-1010	+0.94
<code>retrieval_push k=3</code>	en→ht	full-1010	+1.14 ($p<.001$)
<code>retrieval_push k=3</code>	fr→ht	full-1010	+1.43 ($p<.001$)

en→ht: 41.52 (OFF) → 42.46 ($k=1$) → 42.65 ($k=3$); fr→ht: 41.27 → 42.70. k plateaus at 3, min sim=0.55; into-ht only. It *beats* reasoning (42.65 > 41.93) at lower cost ($1.1\times$ vs $1.7\times$ tokens) but is substitutive with it (+0.01 stacked).

I.5 Grammar-card push (direction-gated)

Targeted morphosyntactic cards (TMA markers, postposed determiners, negation) pushed into the prompt were valuable only where there is no gloss to lean on.

Variant	Dir.	Set	Δ chrF++
grammar_push (ON)	fr→ht	first-200	+0.77
grammar_push (ON)	en→ht	–	–0.30 (redundant)
grammar_push, reason ON	en→ht	full ($n=1005$)	–0.03 (n.s.)
two-pass select (draft×2)	en→ht	full-1012	+0.48 (rejected, $\sim 2.2\times$ cost)

Direction-gated: ON for fr→ht (only evidence), OFF for en→ht (gloss present $\Rightarrow$ redundant; reasoning-on A/B confirms it). Two-pass output selection (base 40.43 | push 40.30 | two-pass 40.77, CI [+0.32, +0.65]) is real but costs $\sim 2.2\times$; default-OFF.

I.6 Reasoning (enable_reasoning)

Enabling the model’s reasoning mode. Its decisive win is repairing a deterministic Creole-passthrough collapse in ht→fr.

Variant	Dir.	Set	Δ chrF++
enable_reasoning	ht→fr	full-1012	+3.02
enable_reasoning (into-ht)	en→ht	full-1012	+1.47
enable_reasoning (into-ht)	fr→ht	full-1012	+0.73

ht→fr: 50.51 $\rightarrow$ 53.54, the recommended ht→fr configuration. Into-ht deltas are over the reasoning-on baseline (additive) and are partly superseded by the retrieval push (which beats reasoning into-ht at lower cost). Reasoning stays the lever out of Creole.

I.7 Prompt levers (few-shot, anti-code-switch)

Held constant, reported separately from the harness (a confound). Static FLORES-dev demonstrations ($k=5$) plus the anti-code-switch rule (“translate every word; do not transliterate the source”) are worth $\sim +1.2$ combined on the wide-gloss stack (en→ht). Dynamic few-shot is the retrieval push above.

I.8 Second-pass / revision

Same-model self-revision after a first draft. revise/post-edit (2nd pass), en→ht, full: –0.40. A second pass over-edits already-correct text and was kept default-OFF. The one second pass that does help is two-pass grammar selection (not self-edit), covered under the grammar-card push section above and likewise rejected on cost.

I.9 Collapse guardrail

A safety net for the degenerate repetition/length-explosion outputs the weak base occasionally produces into Creole. Deterministic normalize_repetition is worth +0.5–1.5 on collapse cases with 0 harm to clean outputs; regenerate_on_collapse (with retrieval, into-ht) fires only on the $\sim 0.3\%$ degenerate outputs and keeps the retry only if it scores better.

J Harness error typology

We group the harness's errors into seven types, each with a full example. The useful question is not whether an error is fixable in principle but where it originates, which we tag on three axes:

R resource: the lexicon or corpus is itself wrong or noisy.

U use of evidence: the evidence was there and correct, but the model picked the wrong sense or over-trusted a bad push.

M base model: its own limits, in target fluency, world knowledge or decoding stability.

In each box the offending span in our output is **marked**, and the **gold form** shown where useful.

src source **gold** gold reference **ours** our output **pushed gloss** what the tool injected

J.1 Hallucination

The output invents content absent from the source such as a fabricated entity, mid-sentence (hallucination).

en→ht • challenge set

src	Indeed, John the Baptizer called Jesus “the Lamb of God that takes away the sin of the world.”
gold	Vrèmanvre, Jan Batis la te rele Jezi “Mouton Bondye a ki rache peche tout moun”. (<i>frontier & Gemma; challenge set has no blind gold</i>)
ours	Vrèmanvre, Madan Viktò te rele Jezi “ti mouton dyè sa ki pran peche mondi”.

M Origin. The verse is rendered faithfully (*Vrèmanvre ... te rele Jezi ... ti mouton*) except that the subject “John the Baptizer” is replaced by an invented person, “Madan Viktò” (Madam Victor); frontier and Gemma both give the correct “Jan Batis la”. A base confabulation, not a gloss item.

J.2 Repetition / decay

The decoder collapses into a fixed n-gram loop or runaway length.

en→ht • base (no harness) • chrF++ 5.0 → 40.8 fixed

src	In fact, it is not easy to find at all even if one knew it existed. Once inside the cave, it is a total isolation.
gold	An fèt, li pa fasil ditou pou jwenn menm si ou ta konnen egzistans li. Depi li anndan gwòt la, se izolman nèt.
ours	Faktè, pa fasil <i>pou w pote yon chans ...</i> (that four-word phrase repeats 96×)

M Origin. A decoding collapse of the weak 8B base into a single repeated phrase; the shipping collapse guardrail (Appendix B) regenerates the sentence and recovers it to chrF++ 40.8.

J.3 Word-sense disambiguation

The correct sense is in the pushed gloss, but the model picks a different offered sense.

en→ht • chrF++ 37.9

src	Plants make oxygen which humans breathe, and they take in carbon-dioxide which humans exhale (that is, breathe out).
gold	Plant yo pwodwi oksijèn imen yo respire, epi absòbe di oksid kabòn imen yo lage (sa vle di degaje).
ours	Plant yo fè oksijèn ki moun yo souffle, epi yo pran gaz kabonik ki moun yo lage (sa vle di ki yo souffle deyò).
pushed gloss	breathe (v): souffle, respire

U R Origin. The pushed gloss lists two senses: “breathe/inhale” (respire) and “blow” (soufle); for “humans breathe oxygen” the model picks the first one, *soufle* (blow), and ignores the better sense. Freeman lists the two senses flat and unranked.

J.4 Push-induced error

The push injects a wrong item and the model faithfully adopts it.

ht→en • challenge set (from-Creole)

src	Jezikri konprann enjistis epi Li gen pouvwa pou pote remèd pou sa.
gold	Jesus Christ understands injustice and He has the power to bring a remedy for it. (<i>unambiguous; frontier and Gemma agree</i>)
ours	The sandal understands injustice and has the power to carry a remedy for that.
pushed gloss	jezikri (n): sandal (usu. leather, without heel strap) (the only sense pushed for a common noun).
R U	Origin. A Freeman entry-bleed defect makes “sandal” the head (and only pushed) sense of the proper noun <i>Jezikri</i> ; the push injects it and the model adopts it over the obvious name. The model copies the retrieved gloss without a critical analysis.

J.5 Calque

A source word or structure is transferred literally, not idiomatically.

en→ht • chrF++ 44.5

src	However, the percentage of XDR-TB in the entire group of people with tuberculosis still seems to be low; 6,000 of the total 330,000 people infected at any particular moment in South Africa.
gold	Sepandan , pousantaj XDR-TB a nan ansanm gwoup moun ki gen tibèkiloz yo sanble fèb; 6,000 moun sou yon total 330,000 moun te enfekte nan yon moman presi an Afrik di Sid.
ours	Men tou , pousantaj XDR-TB nan tout moun ki gen tebe akòz toujou parèt ba; 6,000 nan total 330,000 moun enfekte nan kounye a nan Sid Afrik.
M	Origin. The connector “However” is transferred form-by-form as “Men tou” (“but also”) instead of the idiomatic discourse marker <i>Sepandan</i> . We could also interpret that one as a synonym choice.

J.6 Fluency

Lexically mostly right but non-native: word order, TMA markers, or determiners are off.

en→ht • chrF++ 49.0

src	With Kundalini Yoga the Kundalini energy (enlightenment energy) is awakened through yoga postures, breathing exercises, mantras and visualizations.
gold	Avèk Kundalini Yoga a, enèji Kundalini an (enèji eliminasyon) reveye pa pozisyon yoga yo , egzèsis respirasyon yo , mantra yo ak vizyalizasyon yo .
ours	Ak Yoga Kundalini, enèji Kundalini (enèji eklerasyon) te leve atravè pozisyon yoga, egzèsis soufle, mantra ak vizyalizasyon.
M	Origin. Vocabulary is essentially right, but every postposed plural determiner <i>yo</i> is dropped and a spurious past marker <i>te</i> is inserted where gold is bare-stative. This is a grammatical, not a lexical gap and echoes the base-competence of the model in ht.

J.7 Lexicon-retrieval error (pivot / wrong POS)

The retrieval itself returns wrong data: a pivot false-friend sense, or the wrong POS.

fr→ht • chrF++ 51.6

src	Cette découverte permet également de mieux comprendre l'évolution des plumes chez les oiseaux.
gold	Dekouvèt sa pèmèt tou pou pi byen konprann evolisyon plim yo lakay zwazo yo.
ours	Dekouvèt sa pèmèt tou pou konprann pito evolwasyon pikan nan zwazo.
pushed gloss	plumes: pivot fr→en {feather, quill, nib}; feather → “pm il” (OCR-mangled <i>plim</i>), quill → <i>pikan</i> (thorn): model takes the clean <i>pikan</i> .
R	Origin. The pivot expands “plumes” to three English words and merges their Freeman glosses; the right “feather” gloss is OCR-corrupted to “pm il” while the false friend “quill”→“pikan” (thorn) comes through clean, so the model takes it . This is an error propagated from the processing of the lexical resources.