

The Constitution Speaks Nheengatu: An Open MT System and Reproducible Corpus Pipeline for the Amazonian Língua Geral

Fabrício Carraro¹ Rodolfo Joel Zevallos² Rodrigo Gonçalves de Souza³
Caio Henrique Faustino da Silva⁴ John E. Ortega⁵

¹Universidad Internacional Menéndez Pelayo ²Universitat Pompeu Fabra

³Tribunal Regional Federal da 1ª Região ⁴PUC-Rio ⁵Northeastern University

Abstract

Nheengatu (yrl), the Amazonian Língua Geral, has 6–10 thousand speakers and is co-official in São Gabriel da Cachoeira (Brazil), yet remains poorly served by evaluated Portuguese–Nheengatu MT resources. We present the largest parallel corpus assembled for the language (14.5k Portuguese–Nheengatu pairs), anchored by the 2023 official translation of the Brazilian Federal Constitution; the first MT evaluation set built from transcribed spontaneous speech; and, to our knowledge, the first open-weight MT models dedicated to bidirectional Portuguese–Nheengatu translation.

Three findings organize the paper. First, translation into Nheengatu on held-out constitutional articles plateaus on development data regardless of training length or model size, and much of the remaining error sits in conventional legal terms that the written sources do not attest consistently; a terminology success rate over a frozen lexicon shows that glossary mechanisms at training time and lexical constraints at decoding each raise term realization, the latter at a cost in repetitive output that chrF++ barely registers. Second, a corpus expansion that adds transcribed conversation raises speech chrF++ by 24–27 points while leaving written domains unchanged, a register gap that written-only benchmarks cannot see. Third, a larger NLLB model helps translation into Portuguese but not into Nheengatu on legal or speech text. For Portuguese-to-Nheengatu translation, our final systems reach 31.7–32.8 chrF++ on article-disjoint constitutional text, 64.1–64.5 on the near-domain set (62.1 after removing items with target-side content overlap), and 45.2–47.5 on speech.

1 Introduction

Machine translation nominally covers hundreds of languages, and the most recent effort claims

more than sixteen hundred (Omnilingual MT Team, 2026), which is under a quarter of the world’s seven thousand (Joshi et al., 2020). However, appearing on a list is not the same as being served. The training signal is typically a few thousand sentences of scripture, the evaluation set is a benchmark the language does not appear in, and speakers are rarely asked whether the output is worth reading.

Nheengatu (yrl) is a sharp case of that gap, and an unusual one. Its ancestor is Tupinambá, the coastal language Europeans met in the sixteenth century. Jesuit missions carried it up the Amazon as the language of catechism, where it nativised into the *Língua Geral Amazônica* (Amazonian General Language) and became the dominant lingua franca of a highly multilingual basin, spoken across ethnic lines by Indigenous and non-Indigenous people alike (Cruz, 2011). It was, for roughly two centuries, the language in which the Amazon conducted its business.

The Directorate of 1757 then imposed Portuguese and banned the general language, and nineteenth-century upheavals (the Cabanagem revolt, the rubber boom, the Paraguayan War) displaced or killed much of the speaking population. Nheengatu survived chiefly in the Upper Rio Negro (Figure 1), with an estimated 6–10 thousand speakers today (Cruz, 2011; Quinto, 2025). Since 2002 it has been co-official in São Gabriel da Cachoeira, a mostly Indigenous municipality of 52,000 where UFAM trains schoolteachers in the language.

Nheengatu is a Tupi-Guarani language (Cruz, 2011). It is typologically SVO, agglutinating, postpositional, with person and number marked on the verb by prefixes and plural by the particle *ita*, which is usually written bound in our sources (*mira-ita* ‘people’) (Cruz, 2011; de Alencar, 2024). The morphology is concatenative and moderate rather than polysynthetic.



Figure 1: The Upper Rio Negro. Dark: São Gabriel da Cachoeira, where Nheengatu has been co-official since 2002. Boundaries: IBGE; base map: Natural Earth.

The yrl side of our corpus holds 14,888 distinct forms against 12,393 in the matched Portuguese, with 58% occurring exactly once, so word-level n -grams rarely recur, which is why we report chrF++ and spBLEU (Goyal et al., 2022) rather than word-level BLEU.

Written Nheengatu has no single settled standard: at least four orthographic traditions circulate in our sources (the academic norm of Navarro and Avila that the UD treebank follows, the missionary orthography of the New Testament, the Rio Negro school norm in which the Constitution was translated, and a convention devised for a smartphone localisation project) and D’Angelis (2023) finds none better founded than the others. Our transcribed speech adds a fifth, marking diacritics differently from all of them. A Nheengatu Language Academy was created in 2021; this paper treats no current tradition or institution as uniquely authoritative.

Ignoring diacritics, 13.9% of yrl word types in our training data have two or more realisations, and those types cover 70.3% of running text; the same measurement on the matched Portuguese side returns 1.7% and 31.8%. That the variation is orthographic and not lexical is visible inside a single document: the Constitution alone writes *istadu-ita* 63 times and *istadu-ità* 23, *deputadu-ita* 10 against *deputadu-ità* 30. Corpus-wide, 1,458 sentences write the plural particle bound, 298 write it free, and 48 do both. A system must emit some orthography, and whichever it emits will be unfamiliar to part of its readership; which norm to produce is therefore a community decision, and our contribution is to build the mechanism that can serve whichever decision is taken.

That co-officiality has no technological counterpart: as of August 2026 the language is absent

from NLLB-200, FLORES+, MADLAD, Google Translate and the 1,600-language release of [Omnilingual MT Team \(2026\)](#), and to our knowledge it has no public MT benchmark. The multilingual open-weight release OmniTranslate lists yrl among its languages but reports no yrl-specific evaluation (MihaiPopa-1, 2026). The prior published Portuguese–Nheengatu MT results we discuss come from closed industrial research (Gonçalves et al., 2025).

Two circumstances make this the moment to build one. In 2023 Brazil’s Federal Supreme Court and National Council of Justice published the Federal Constitution in Nheengatu, translated over two years by fifteen Indigenous translators under editorial review. It is a contemporary, long, non-religious text of a kind that projects at this resource level almost never have. And enough other material exists (dictionaries, a treebank, school books, transcribed conversation) that a usable corpus can be assembled if provenance is tracked source by source.

Contributions. (1) **Corpus** (§3): 14.5k pt–yrl aligned pairs with source provenance and orthographic metadata, twice the closed corpus of Gonçalves et al. (2025), together with article-disjoint legal, pair-deduplicated near-domain, and speech evaluation sets with speakers and sessions held out from training. (2) **Models** (§5.6): to our knowledge, the first open-weight MT models dedicated to bidirectional Portuguese–Nheengatu translation. (3) **Terminology, diagnosed and treated** (§5.2): a terminology success rate (TSR: the share of glossary terms present in the source that the output realizes), scored over a lexicon frozen before the terminology experiments, identifies unstable conventional terminology as an important contributor to the legal plateau; annotation and token-pair augmentation improve term realization; constrained decoding raises surface-form realization at a $3.2\text{--}4.8\times$ rise in degeneration (repetitive, runaway output), a cost chrF++ barely registers. (4) **Register vs. model size** (§5.3, §5.4): the corpus expansion strongly improves speech while leaving written-domain scores nearly unchanged; within the tested NLLB family, scaling improves yrl→pt and near-domain pt→yrl but not article-disjoint legal or speech pt→yrl in our runs.

2 Related Work

MT for yrl (closed). IBM Research Brasil fine-tuned NLLB-600M on $\sim 6.7k$ manually cleaned pairs, reporting 69.1/66.9 (pt $\rightarrow$ yrl / yrl $\rightarrow$ pt) on a 233-segment test of short dictionary sentences (Gonçalves et al., 2025). Those figures are *sentence-level* chrF, with large reported standard deviations of ± 33.8 and ± 34.6 ; they are not corpus-level chrF++, and they are computed on a private test set; we therefore treat them as context, not as a baseline we can measure a gap against (§5.6). Their corpus and models are unpublished, and the same group’s human evaluation of those translators found automatic metrics and speaker-judged usefulness to diverge substantially (Pinhanez et al., 2024), one reason we treat our own chrF++ gains as necessary but insufficient evidence.

Low-resource and Indigenous MT. No shared task has ever included Nheengatu: neither WMT (Kocmi et al., 2024a, 2025) nor any of the five AmericasNLP editions covers it, and it is absent from the 202 languages of NLLB-200, whose only Tupian language is Guaraní. Because no benchmark contains this language, there is no system we can rank against, and we make no ranking claim anywhere in this paper; what counts as a meaningful difference in any case shifts with direction, domain and system proximity (Kocmi et al., 2024b).

Choice of base model. Five years of AmericasNLP findings converge on one recipe: fine-tuned NLLB (NLLB Team et al., 2022, 2024) remains the strongest system for translation *into* Indigenous languages (Mager et al., 2021; Ebrahimi et al., 2023, 2024; De Gibert et al., 2025; Stap and Araabi, 2023), to the point where the 2025 edition reports a fine-tuned NLLB baseline that a single team beat, for a single language, by 0.82 chrF++ (De Gibert et al., 2025). Grammar-book prompting works mainly for the reverse direction and for languages with richer descriptive material (Tanzer et al., 2024; Aycock et al., 2025; Pei et al., 2025; Zhang et al., 2024a,b), and post-correction pipelines inherit that asymmetry (Hus et al., 2025); in-context and modular adaptation of large models assume retrieval pools or per-language data far beyond what yrl has (Moslem et al., 2023; Xu et al., 2025). Our own LLM post-editing experiment is a documented null (§5.7), and our same-corpus architecture control is mixed once checkpoint selection is equalized (§5.5).

source	pairs	yrl w.	w/s	register
Constitution 2023	2,031	42,262	20.8	legal
Avila 2021 (dict.)	2,745	24,437	8.9	dictionary ex.
UD treebank	2,484	19,270	7.8	mixed
LEETRA (4 books)	526	8,284	15.7	school
Melgueiro 2022	1,870	8,128	4.3	dictionary ex.
Gurgel (compil.)	1,562	7,538	4.8	teaching
Refubium (speech)	1,156	4,474	3.9	conversation
Tycho Brahe	441	4,269	9.7	historical
Navarro (course)	187	966	5.2	teaching
Catechism 1944	23	276	12.0	religious
pt-yrl training	13,025	119,904	9.2	
<i>held out evaluation sets</i>				
legal dev / test	233 / 218	4,421 / 3,475	17.5	legal article-disjoint
near-dom. dev / test	372 / 371	2,834 / 2,701	7.4	dictionary
speech dev / test	116 / 115	487 / 537	4.4	conversation

Table 1: Corpus composition, ordered by yrl words. “yrl w.” is running words on the Nheengatu side; “w/s” is the mean number of words per aligned segment; speech units are pause-delimited turns rather than sentences. Religious text is **0.2%** of the corpus (23 pairs), where for most of the world’s languages scripture is the only large digital resource; our anchor is instead a contemporary legal text translated under editorial review.

3 The Corpus

3.1 Constitution pt-yrl (anchor)

In 2023 the Brazilian Supreme Court and the National Justice Council published the Federal Constitution in Nheengatu (Supremo Tribunal Federal and Conselho Nacional de Justiça, 2023), translated over two years by a team of Indigenous translators of the Upper Rio Negro and the Middle Tapajós. It is to our knowledge the largest modern text in the language, and the only one that is simultaneously long, contemporary, and translated by speakers under editorial review. It is also awkward as a corpus: a PDF of a printed book with no sentence alignment, facing an official Portuguese edition that has its own amendment history.

We align the two by *document-structure identity* rather than by length or similarity. Both editions carry the same legal skeleton (article, paragraph, *inciso* and *alínea*: the numbered item and lettered sub-item of Brazilian legal drafting), so we parse each side with a state machine that validates identifier sequences as it walks the document (article 5 must be followed by article 6; item XII by item XIII), and pair units whose full identifier path matches. Structure-based alignment is what makes the result auditable: every pair carries the legal address that produced it, so any error is locatable rather than statistical. Alignment yields 2,496 pairs. The final filtered corpus retains 2,482: 2,031 for training, 233 for development and 218 for testing (Table 1).

We verify the Portuguese side against a third source, the Planalto reference edition (Appendix A). The exercise also surfaced errata in the printed Nheengatu book: mislabeled items, cascade renumberings, and an entire constitutional-amendment block absent from the translation. We detected these precisely because the structural addresses disagreed with the text. Renumberings we repaired by content; six pairs with genuine content errors we excluded from the evaluation splits instead of silently correcting them. Development and test splits hold out *whole articles* (19 and 19), not sentences, so that no article contributes to both training and evaluation, with overlap verified to be zero on both sides.

3.2 The aggregate corpus and its versions

Every pair in Table 1 carries its provenance and the orthographic tradition it was written in: a pair’s spelling is only interpretable if one knows which of the competing norms produced it, and any later normalization decision must be reversible. Three corpus versions are named in what follows. v1 is the aggregate as first assembled. v2 applies an audited cleaning pass removing dictionary apparatus, ellipsis markers and typographic ALLCAPS; it changes 170 evaluation pairs, so numbers from the v1 series are never point-compared with later ones. v3 adds 3,408 audited training pairs: (a) **transcribed spontaneous speech**; (b) **dictionary example sentences** (audited against page sources: zero cross-pairings); (c) **school books** from two previously uncovered dialect areas, where length-based alignment *failed* under free translation (one book places translations *before* the source stories) and was replaced by semantic realignment with literal anti-hallucination validation (99.6% pass). Speech evaluation holds out an entire session group: 116/115 content-deduplicated, pause-delimited turns from its two speakers across four elicitation task types, disjoint from the nine speakers in training; the speech test set is single-use. The near-domain sets are exact-pair-deduplicated only (not source- or document-disjoint) and the Limitations quantify what that leaves. Audits of 200 speech pairs and 240 dictionary-example pairs approve 91.5% and 86.7%, respectively. The residual issue classes account for 1.0% and 6.7% of the audited samples. The sampling procedure was not preserved, so these figures do not establish defect rates for the unaudited remainders. These are not one minus the approval rate, because the two non-

ok tails have different fates; Appendix A gives the categories and the derivation. Neither residual is mispairing, which is tolerable in training data and would not be in evaluation data.

Training additionally includes 925 FR–yrl pairs (a Nheengatu translation of *Le Petit Prince*, 12,467 yrl words): the Nheengatu side is human, the source side French. They are present in every run reported here. A Portuguese view of the same targets, translated from the French by a frontier LLM (Claude Fable 5), is a separate experiment, reported and rejected in §5.2.

4 Experimental Setup

We fine-tune NLLB-200-distilled at 600M and 1.3B (NLLB Team et al., 2022, 2024), using the 600M model for replicated configuration comparisons and the 1.3B model for single-seed scale comparisons. Nheengatu is not explicitly supported by NLLB-200 and has no dedicated language token, so we add yrl_Latn and initialize it from the Guarani grn_Latn embedding. Training is bidirectional, so a single model serves both directions.

Training glossary. Two mechanisms are trained jointly. Inline annotation follows target-lemma annotation (Bergmanis and Pinnis, 2021): a source term is followed by its yrl equivalent in a delimited slot (term | yrl |), applied to 30% of matched term occurrences in training and, at *hinted* inference, to every source-side glossary match (the reference is never consulted), capped at two matches per segment. Token pairs follow the lexical-augmentation recipe (Jones et al., 2023): the glossary’s 1,027 surface forms injected as standalone training pairs. Both draw on the same audited glossary of 904 entries, mined by Dice association and adjudicated against source evidence. The glossary is built only from training pairs and external lexical resources; evaluation references are never used to construct or adjudicate it.

Scoring lexicon and Terminology Success Rate. TSR follows the terminology success rate of the WMT terminology task (Semenov et al., 2023): the proportion of scoring-lexicon term occurrences in a source sentence whose yrl equivalent (or a listed synonym) appears in the output. The scoring lexicon contains the training glossary. We freeze scoring lexicon v1 *before* the terminology experiments by taking the union of the audited training glossary and the base bilingual lexicon. Every TSR value

	legal (const)		near-dom.	
	pt→yrl	yrl→pt	pt→yrl	yrl→pt
600M (3 seeds)	31.25 ± 0.44	35.45 ± 0.31	64.12 ± 0.38	56.12 ± 0.41
1.3B (1 seed)	32.03	37.76	66.37	58.22

Table 2: Phase-1 evaluation (v1 series): chrF++ for the 600M (mean $\pm$ sd over three seeds) and the 1.3B (one run). The 1.3B differences are descriptive: a single run has no seed spread of its own, so the 600M’s $\pm$ sd does not standardize them.

in this paper uses the same frozen scoring lexicon. A hit counts exact, lemma-table, accent-insensitive and fuzzy matches alike, the last at 0.85 and only for forms of five characters or more. Under this protocol the legal dev set carries 198 term occurrences and the legal test set 232. TSR remains conservative: §6 shows by per-case verification that it *understates* our mechanisms, because inflected, re-spelled and fused realizations still count as misses.

Metrics. We report chrF++ as the primary metric and do not use a learned metric as primary evidence, because COMET, CometKiwi and MetricX have not been validated for yrl. We attach no usability label to any absolute chrF++ value.

Decision discipline. Configuration decisions use full-dev evaluation and require three seeds; the noise ruler on the legal dev set is ± 0.35 chrF++. We adopted the three-seed rule after two single-seed results failed to replicate, and we report both. Test sets are pre-specified and decoded once per model; §5.6 states which are pristine and which are reused holdouts. A 20k-step 600M run costs ≈ 1.3 GPU-hours, which is what makes three seeds per configuration affordable. Every configuration decision here is reproducible on a single GPU, and only the 1.3B confirmations need a full node. All $\pm$ values are sample standard deviations over training seeds.

5 Results

5.1 Phase 1: configuration and scale effects

Phase 1 settles two configuration choices by ablation: related-language mixing and orthography. Table 2 reports the resulting Phase-1 evaluation on the legal and near-domain holdouts. Mixing Guarani–Spanish data, the obvious transfer candidate given that our new token is initialized from Guarani, costs 0.26 chrF++ and is discarded: the embedding is

training mechanism	unhinted	hinted
none (baseline)	51.5	not run
token pairs only	55.1	not run
annotation only	51.0	63.6
annotation + token pairs	54.5	66.2± 2.7

Table 3: TSR on legal development data (198 term occurrences): training mechanism crossed with decode-time hints (v2, 600M). We did not evaluate hints on models trained without annotation; their response to unseen markup would also test robustness to the input format. Single-seed values, except the hinted combination (three-seed mean).

a starting point, not a related-language signal the model can exploit.

We expected orthographic unification to help, and it did not. Only 29% of the Constitution’s *tokens* (5% of its types) appear, by exact surface match, in the vocabulary of the largest modern dictionary (Avila, 2021); ignoring accents raises these to 46% and 8%, which is itself a measure of how much of the divergence is diacritic. Unification nonetheless moved every evaluation set by less than seed noise, with the legal set landing 0.3 chrF++ lower, and it did so with the evaluation references themselves favouring it. We retain a light normalization (punctuation, casing, whitespace) and preserve the original spelling of every pair alongside it. The result is specific to the recipes and data tested here.

Across 8k and 20k training steps and across both model sizes, legal pt→yrl sits at ≈ 29 chrF++ on dev. Essentially any configuration lands in the same place, while near-domain translation and both yrl→pt directions respond normally to training length and model size. The plateau suggests a data bottleneck specific to article-disjoint legal pt→yrl. §6 identifies unstable conventional terminology as one important contributor, but the present experiments do not establish that it is the sole cause.

5.2 Phase 2: terminology-control ablations

The initial mechanism ablations in Table 3 use 600M models on legal development data for pt→yrl with corpus v2. *Hinted* means glossary terms are supplied inline at decoding; *unhinted* means they are not. Table 3 distinguishes the measured training-only comparison from the within-model effect of hints. Relative to the no-mechanism baseline (28.91 chrF++, 51.5 TSR), the combined mechanism alone, decoded without hints, adds 3.0 pp TSR (54.5, single seed) and 0.99 ± 0.17

chrF++ over three seeds. On the one seed decoded in both conditions, supplying hints adds a further 8.6 pp (54.5 to 63.1); across three seeds the hinted system reaches 66.2 ± 2.7 (range 63.1–68.2), 14.7 pp above the unhinted baseline, and brings chrF++ to 1.24 ± 0.32 above it. The no-annotation hinted baseline was not evaluated, so the interaction between annotation training and hints is not estimated. Annotation alone changes chrF++ by 0.29, within seed variation, and token pairs alone add 3.6 pp unhinted TSR. Hints apply to 110 of 233 segments, so the aggregate effect is bounded by glossary coverage.

Negative results: the LLM-translated Portuguese view of the 925 literary pairs reduces chrF++ by 0.66 in the unhinted condition and by 1.5 in the hinted condition. Because the Portuguese view is added rather than substituted, the experiment jointly changes duplication, source-language weighting, register, and orthographic composition.

In an exploratory failure-targeted distillation experiment on corpus v3, relative to the annotation-plus-token-pair configuration, we identify 43 terms that fail on legal development data, write 10 legal-register Portuguese sentences for each (leak-guarded against every evaluation set) and translate them with the term forced by constrained decoding, keeping the 233 that survive a length filter. These yield at 600M (2 seeds) **+0.93 chrF++ on legal dev** (30.91 ± 0.25) and **+4.2pp unhinted / +7.2pp hinted TSR** ($60.4/72.2$), with hint chrF++ turning neutral: the glossary migrates into the parameters. A single 1.3B run with the same distilled pairs shows a similar unhinted TSR gain (+4.6pp, 54.5→59.1; single seed) while *halving* the model’s dependence on the hint (hint effect +8.1pp→+4.5pp), though the 600M’s chrF++ gain does not transfer (+0.03): at 1.3B the distilled glossary shows up purely as terminology, invisible to chrF++. Term selection is dev-guided (declared).

The mechanisms address different parts of the problem: token pairs expose the model to a term and its spelling, while annotation trains it to condition on a supplied target form and place it in context, which is what makes hinting possible at all (51.0 unhinted to 63.6 hinted). Constrained decoding raises the realization rate at inference, but it is lexical enforcement rather than a guarantee of correct terminology.

eval	v2 model	v3 model	Δ
speech pt→yrl	23.06 [†]	46.77	+23.7
speech yrl→pt	22.82 [†]	49.81	+27.0
legal pt→yrl	29.90 ± 0.17	29.98 ± 0.13	+0.1
near-dom. pt→yrl	60.70 ± 1.01	59.99 ± 0.64	−0.7

Table 4: Effect of the v2→v3 expansion (3,408 pairs, 1,156 of them speech) on development chrF++. [†]The v2 model saw no speech in training, so the speech rows measure the whole expansion, not the speech subset alone; those two rows are single runs without terminology mechanisms, while the written rows are three-seed means of the shared configuration. The expansion also raises unhinted TSR by 3.0 pp, an effect not attributed to any single source.

5.3 Phase 2: the speech-register gap

The size of the gap in Table 4 is worth dwelling on, because it is easy to misread as a data-volume result. A model trained on written Nheengatu (legal text, dictionary sentences, school materials) is not merely worse at spontaneous conversation; at 23 chrF++ its similarity to the references falls far below the 30 it reaches on the hardest written set and the 60 on the easiest. Moving from v2 to v3 (an expansion of 3,408 pairs, 1,156 of them speech, under 9% of the resulting 13k-pair training set) improves speech chrF++ by 23.7 to 27.0 points while changing written-domain scores by less than one. The evaluation is speaker-disjoint and session-disjoint, which reduces the risk of speaker memorization. The concentration of the gain on the speech evaluation is consistent with a matched-register effect, but the comparison does not isolate the speech subset from the other v3 additions. We report this because projects on languages that live primarily in speech are routinely evaluated only on written benchmarks, where a 23-point blind spot is invisible.

What the speech errors look like. The v2 model’s speech outputs were not retained, so we cannot place them beside v3; we characterize the v3 model of Table 4 instead, on its 116 development turns. The lowest-scoring outputs are one- and two-word turns whose reference depends on discourse context the turn does not carry: *Asu* (‘I went’) rendered as *vou* (‘I go’), *amuyatimana* (‘I turned’) as *ajunto*. Mid-range outputs are adequate paraphrases that chrF++ scores in the forties: reference *mas da nossa região ou de outra*, system *mas da nossa terra ou outro*. Into Nheengatu the dominant residual is orthographic: *siía pixana ta uiku tatá rendá ruaki* against the reference *siyawa*

pixana uiku tata rēda ruaki differs mainly in accent and nasal marking and scores 42. Two adequacy errors recur: a vocative name rendered as a kinship noun (*Dadá* → *Papai*, ‘Dad’), and, in the larger and terminology-augmented models we inspected on the same turns, the reportative *disque* inserted into Portuguese where the source carries no evidential marker.

Register-control codes were validated and then rejected for the final system. A [falə] tag prepended to conversational sources buys +2.34 chrF++ on speech, a real gain above the noise ruler, but costs a real −0.74 on legal text, and the tagged model collapses to 26 on speech when the tag is absent, meaning it has learned to depend on a signal that a deployed system would have to infer. For a single general-purpose model serving mixed traffic, we do not adopt this trade-off. For a dedicated conversational checkpoint, the configuration is promising but requires speaker and user evaluation. Terminology behaviour is unaffected either way, which is itself informative: register control and terminology control behave as largely separate axes in this experiment.

5.4 Phase 2: scale at fixed configuration

Phase 1 established that model size does not move legal pt→yrl. With the optimized configuration and the v3 data in place, we can ask a sharper question: is that a property of the legal domain, or of generating Nheengatu at all? Holding the configuration fixed and moving only from 600M to 1.3B, legal pt→yrl lands at 29.67, just below the 600M three-seed spread (29.89–30.12), and spontaneous speech at 46.17, inside the speech spread (45.79–46.70), while translation *into Portuguese* gains 1.3, 3.0 and 2.5 chrF++ on legal, near-domain and spontaneous speech respectively, and near-domain into-yrl gains 3.1.

Within the tested NLLB family, scaling helps all yrl→pt settings and the easier near-domain pt→yrl setting, but does not resolve the harder article-disjoint legal or spontaneous-speech pt→yrl cases in our runs. Because the 1.3B condition uses a single seed, these differences are descriptive.

The direction gap is consistent with NLLB’s own ablations, in which translation into English exceeds translation out of English by a larger margin for very-low-resource directions (NLLB Team et al., 2022). Nheengatu is not explicitly supported by NLLB-200 and has no dedicated pretrained language token.

Hinting continues to help at 1.3B on domain text (+0.55 legal) but costs −0.95 on conversational input, where dictionary-form hints reduce performance. The operational rule we adopt is to hint domain text but not conversation.

5.5 A same-corpus architecture control

We pre-specified a same-corpus control before measuring: full fine-tuning of Qwen3.5-2B (Qwen Team, 2026), a 2B instruction-tuned decoder-only model with a hybrid-attention stack, against our 600M seq2seq, three seeds each, on the development sets. The two sides were not selected alike: the seq2seq kept the checkpoint with the best development chrF++, whereas the decoder-only kept the one with the lowest development loss, which is step 1,000 in all three seeds. Under loss-based selection the seq2seq leads all six cells (Table 7). Because the criteria differ, after submission we also scored the decoder-only’s late checkpoints (steps 11–12k, chosen with knowledge of the development scores as a proxy for chrF++-based selection) with the same decoder. The picture changes: the seq2seq still wins conclusively on article-disjoint legal pt→yrl (+1.90 chrF++ against a threshold of 1.06 under the pre-specified rule), the decoder-only wins both near-domain cells (−1.50 and −3.23), and the other three cells are inconclusive. Our declared hypothesis, that the decoder-only would compete into Portuguese, is therefore supported on near-domain yrl→pt once selection is equalized, and not on legal text. Two cautions apply: the late checkpoints are a post-hoc choice, and the near-domain set is where content overlap with training is highest (Limitations), which favours whichever model reproduces training material more readily. This compares *systems*, not architectures (pretraining, size, instruction tuning and checkpoint selection differ together), with one model per side and on development data only, so it does not generalize to decoder-only models as a family. Appendix A gives both selections, the six-cell table and the hyperparameters.

5.6 Phase 2: final evaluation

Table 5 reports it, for two systems: the **flagship** (1.3B, light punctuation/casing/whitespace normalization + annotation + token-pair augmentation, 20k steps) and **avg**, the weight average of steps 29k and 30k of a 30k-step extension of that same run. Note also that the two held-out article sets differ in difficulty by ≈2–3 chrF++ (legal test is the

system	legal		near-domain		spontaneous speech	
	pt→yrl	yrl→pt	pt→yrl	yrl→pt	pt→yrl	yrl→pt
flagship	31.70	37.39	64.11	57.10	45.24	49.67
+ hints	31.13	—	64.42	—	—	—
avg	32.78	38.14	64.54[†]	58.76	47.45	49.37
avg, v1 refs	32.74	38.15	62.55	56.92	—	—

Table 5: Phase-2 final test, decoded once. For pt→yrl, paired bootstrap intervals do not resolve the difference between our two systems on legal text (95% CI $[-0.06, +2.26]$) or near-domain text ($[-0.48, +1.44]$); the speech interval excludes zero ($[+0.46, +4.06]$). The same holds against Phase 1 on identical v1 references (32.74 vs. 32.03, $[-0.37, +1.79]$, $p = 0.20$). The 0.71 gap against Phase 1 on v1 references is not resolved by the paired-bootstrap interval on this 218-segment set. [†]Removing the identified target-side content-overlap items lowers this to 62.09; we report both, and the Limitations give the overlap measurement.

system	language	test set	→Indig.	Indig.→
ANLP-25 base	Guarani	XNLI	35.7	35.9
ANLP-25 base	Quechua	XNLI	31.9	37.2
ANLP-25 base	Nahuatl	XNLI	22.4	26.4
ANLP-25 base	Otomí	XNLI	12.8	20.8
ours 1.3B	Nheengatu	legal (article-disjoint)	32.8	38.1
ours 1.3B	Nheengatu	speech	47.5	49.4
ours 1.3B	Nheengatu	near-dom.	64.5	58.8

Table 6: Context from AmericasNLP, not a ranking: no cell is a like-for-like comparison. All values are chrF++ from fine-tuned NLLB-200-distilled-1.3B models with an extended embedding matrix; our rows are the *avg* system of Table 5. The AmericasNLP pairs use Spanish and translated XNLI test sets; ours use Portuguese and single-step translations (field translation on the speech side, see Limitations).

easier draw), so dev and test numbers are reported separately throughout and never pooled; the Phase-1→Phase-2 comparisons below use the same evaluation split and reference series. Trivial baselines fix the scale: copying the Portuguese source scores 13.8 / 12.2 / 9.3 chrF++ on the legal, near-domain and speech test sets, and shuffled references 13.1 / 9.7 / 11.7. Table 6 fixes the other end: the spread between our own three evaluation sets is larger than the spread between published systems for comparable languages, which is itself the point.

The legal and near-domain sets were first evaluated in Phase 1 and reused in this evaluation, so we treat them as reused holdouts rather than pristine single-use test sets. Each Phase-2 model–test combination was decoded once, with no test-based checkpoint selection. The speech test set remained untouched until this pass. Before decoding we fixed, in writing, which checkpoint would be eval-

uated, which metric would decide, and what each possible outcome would mean.

First, the ranking flipped between dev and test. On dev the flagship checkpoint beat the checkpoint-averaged system; on test the average won the headline metric by more than a point on the legal set. Had we selected on dev we would have promoted the system test ranked second; had we selected on test we would have been fitting it. We report both, which quantifies the selection risk a 218-segment set carries.

Second, cross-series scoring punishes cleaner models: scored against the *original* v1 references our v3 systems reach 62.6 on near-domain against 66.4 for the v1-trained system, because a model trained not to emit the artefacts those references still contain is penalised for their absence. We therefore never compare across reference series.

Third, the closed baseline (Gonçalves et al., 2025) cannot be compared to ours directly. Rescoring our outputs in *their* protocol gives 64.48 pt→yrl and 60.2 yrl→pt on near-domain against their 69.1 and 66.9. What matters more is that their test set is a random split of the corpus the model was trained on (6,699 training pairs, 233 test pairs, randomly divided); the same paper splits its other language by source instead, and reports 40.2 there. A random split of the training corpus and a held-out test of unseen legal prose measure different things, and only the second probes unseen legal content.

The test also reproduces the terminology dissociation. Without any inference-time intervention, the Phase-2 system scores 47.8 TSR against the Phase-1 system’s 44.0 on the same 232 term occurrences (**+3.8pp**): data and training alone internalised part of the glossary. Adding hints then moves TSR from 47.8 to 57.3 (+9.5pp) while costing 0.57 chrF++, the same dissociation observed on development data, reproduced under the pre-specified Phase-2 evaluation. The averaged system, which obtains the highest chrF++ of our two final systems, simultaneously obtains the lower terminology score, at 44.4 TSR.

5.7 Decoding without a validated neural metric

Because no learned quality metric is validated for yrl, we do not use neural-metric utility as primary evidence for reranking. We instead evaluate search variants, string-based MBR, terminology reranking, and lexical constraints. Five are null. Three are measured on the near-domain dev set against a

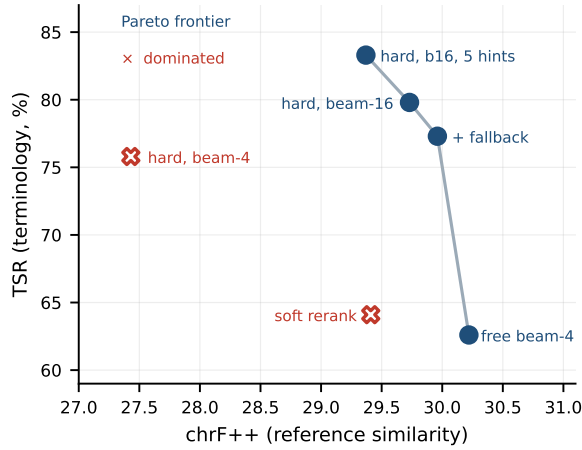


Figure 2: Terminology realization versus chrF++ on hinted legal development data. Free beam-4 gives 30.22 chrF++ at 62.6 TSR; hard constraints with a wide beam and five hints reach **83.3** TSR at 29.37. Two settings are *dominated* in both dimensions at once: soft reranking (29.41/64.1) and hard constraints at beam-4 (27.43/75.8), where narrow constrained search substantially reduces reference similarity and a wider beam recovers part of the loss, as the dynamic-beam-allocation literature predicts (Post and Vilar, 2018). The decoder realizes 164 of 165 requested surface-form insertions, or 99.4%. This measures surface inclusion rather than semantic, morphological, or syntactic correctness. The wide-beam gain is hint-conditional: the argmax drops only when hints are present.

beam-4 reference of 63.1 chrF++: sampled chrF-MBR (61.9–62.6, even with the beam candidate in the pool); MBR reranking of the beam’s own n-best (damages the argmax); and wider beams, which saturate unhinted. The other two are conservative LLM post-editing of the Portuguese side (−0.2 chrF++; no improvement in reference similarity) and soft terminology reranking, dominated on the legal-dev frontier below. One frontier does validate (Figure 2).

A cost chrF++ almost hides: constraints multiply degeneration. Repetition detection over the same 233 outputs (punctuation-normalized n -grams; **zero false positives on the 233 human references**, and no source sentence contains list markers) flags 1.7% of free beam-4 outputs and 1.3% of hinted ones, against 6.0%, 7.7% and 8.2% for the three constrained settings (paired uncorrected exact two-sided McNemar $p = .031/.003/.0015$; the ratio holds at 3.2–4.8 $\times$ across four detector thresholds). The failure localizes to the constraints themselves: constrained runs consume the *plain* source (terms enter only through `force_words_ids`), and

among the 110 sentences that actually receive constraints degeneration rises from 3.6% to 16.4%, while constraint-free sentences under the same decoder stay at 0.8%. Every flagged output is a genuine loop on inspection (median length ratio 3.2, up to 17.8 $\times$ the reference, sometimes fabricating enumerations absent from the source).

chrF++ does penalize these sentences (−7.2 each against the same sentences decoded freely), but its recall-weighted $F_{\beta=2}$ makes appended repetition cheap: a hypothesis consisting of the reference *tripled* still scores 71.0. Corpus-level the whole pathology is therefore worth only −0.59.

A repetition-triggered fallback (detect the loop, re-decode that sentence without constraints) buys it back. It fires on 19 of 233 sentences and gives the fallback point of Figure 2: 29.96 chrF++ (within 0.26 of unconstrained decoding), degeneration down from 8.2% to 0.9%, and TSR at 77.3. That last figure is the honest cost: the fallback trades −6.0pp of terminology for +0.59 chrF++, because the rescued sentences lose their forced terms by construction. Whether that trade-off suits a community writing assistant must be settled by speaker and user evaluation.

6 Analysis

chrF++ $\times$ TSR dissociation. Scale buys chrF++ without terminology (test TSR flat at 600M $\rightarrow$ 1.3B in Phase 1); annotation + token-pair augmentation buys both; extended training buys chrF++ *and loses* terminology (the avg system). Per-case verification of all 34 hint-induced term flips shows 24/26 gains syntactically integrated and only 1 real regression: the matcher-based TSR *understates* the mechanism.

Register-conditional orthography switching: trained on multiple traditions, the model in qualitative inspection sometimes follows the orthographic tradition associated with the input register; we do not quantify this effect, and we propose per-norm control codes as the community-governed mechanism.

Terminology errors on article-disjoint legal text. We inspect both model sizes on the same 20 legal-development segments, 40 outputs in all. Outputs generally preserve grammatical frames, legal citations such as *art. 89, VII*, and many numerical expressions, while much of the residual error involves conventional legal terms such as *civil registry*, *death certificate*, and *cruel* (from *penas*

cruéis, ‘cruel punishments’, art. 5, XLVII). The 90 missed term occurrences correspond to 60 distinct glossary forms; only eight occur at most twice in training, and the median training count is 10. Simple token rarity is therefore not the only explanation. Several terms have competing or inconsistent realizations across the written sources we examined. We treat this instability as an important bottleneck, but the present sample does not establish that terminology accounts for the entire legal-performance plateau.

Two further observations from the same samples. First, in $\geq 3/20$ legal samples the system is *more faithful to the source than the human reference*: it renders “sixty” where the reference prints “65”, and “two rounds” where the reference has “morning or afternoon”. chrF++ against such references has a ceiling well below 100, which is why our speaker protocol asks evaluators to judge against the *source*. Second, the hint mechanism’s own failure modes are narrow but real: noun hints pasted into verbal slots, and a digit juxtaposed with the hinted native numeral (“*10 ára-ita mukūi pu aára*”, literally “10 days two hands”); conversely, in two cases a hint *cured* a degeneration loop.

7 Discussion

Article-disjoint legal pt→yrl plateaus across the tested training lengths and model sizes. Our error analysis points to unstable conventional terminology as an important contributor: the largest modern yrl dictionary (Avila, 2021) provides a usable match for only one of 54 outstanding legal terms, and the tested increase from 600M to 1.3B did not resolve terms that are inconsistently attested in the written resources we examined.

The analysis has a practical implication. For languages whose written resources do not consistently attest specialized terminology, a community-governed control interface can complement the model: a revisable glossary, mechanisms that expose the model to it during training, and an optional lexical-enforcement mode at inference time. The hint budget matters as well: allowing four inline hints per sentence at inference, instead of the two used in training, raises TSR by 1.9 ± 0.3 pp on the v3 600M models (three seeds) at no chrF++ cost, because sentences that match more than two glossary terms otherwise lose the rest at the cap. The decoding mode increases specified surface-form realization but does not guarantee semantically, mor-

phologically, or syntactically correct use. The components are independently replaceable, allowing the pipeline to incorporate an authoritative terminology list if and when one is established.

Within the tested NLLB family, increasing model size is consistently helpful for yrl→pt and also improves near-domain pt→yrl, but it does not improve article-disjoint legal or speech pt→yrl in our runs. A larger checkpoint is therefore promising for reading support, whereas writing support in the harder registers is likely to require better matched data, terminology governance, and user evaluation in addition to model size.

Finally, the three dissociations we report (chrF++ against terminology, chrF++ against degeneration, dev ranking against test ranking) all point the same way: for languages without a validated learned metric, a single automatic number gives a projection of quality along one axis, unstable at this sample size, and never a summary of it. The next required instrument is speaker-grounded evaluation of adequacy, naturalness, terminology, and usefulness. The planned protocol is described in the Ethics Statement, but no speaker-evaluation result is included in the present version.

Limitations

Limited architectural coverage. Our principal systems are NLLB-200 fine-tunes with an added language token. §5.5 reports a same-corpus control against a 2B decoder-only system, but we did not train an alternative encoder-decoder base such as mBART or M2M-100, and the control itself covers one model and one scale per side, on development data only.

- Small evaluation sets (218–372 per set; speech 115–116). At this size differences of about a point were not resolved by our bootstrap intervals (Table 5).
- No speaker evaluation is included in this version.
- chrF++ penalizes valid cross-orthography outputs.
- TSR synonym coverage is incomplete, and shown to understate.
- The 1.3B models are single-seed.
- Failure-targeted synthetic term selection is development-guided, and no independent test result is included. We therefore treat the distillation result as exploratory.

- Brazilian text-and-data-mining (TDM) legality for third-party research use lacks an explicit exception. Our posture is non-commercial research; one third-party layer is redistributed with its author’s written permission and the others remain pending.
- Domain status of the legal evaluation. The legal development and test sets hold out complete constitutional articles, but they come from the same translated Constitution as 2,031 training pairs. They are article-disjoint and segment-disjoint, not domain-disjoint or document-disjoint.
- Terminology governance. The glossary is source-derived and researcher-adjudicated rather than community-authorized. Its forms should be treated as provisional alternatives for experimentation, not as an official Nheengatu terminology standard. Its reach is also bounded: in the legal test set, 8.7% of content tokens (176 word types) never occur in training (by surface form) and none has a glossary entry, so no glossary-based mechanism can act on them.

Content overlap in the near-domain test set.

Exact-pair deduplication is not sufficient when a language has few sources. Our near-domain test contains no leaked pair (zero identical source sentences, zero identical targets, zero identical pairs against training), but 23 of its 371 targets share an 8-gram with training data, because the same traditional narratives appear in both the dictionary and the treebank, in different renderings. Those 23 items score 80.5 against the set’s 64.5, and removing them lowers the set to **62.1**. We report both: 64.5 for the declared set and 62.1 as the conservative reading. The legal and speech sets are practically unaffected (1/218 and 0/115 targets, respectively). The development sets show the same pattern (3/233 legal, 25/372 near-domain, 0/116 speech), which matters because the mechanism and decoding results are measured there. This overlap reflects a structural challenge in low-resource corpus construction, and content-level n -gram overlap should be reported alongside exact pair-level deduplication. On the Portuguese source side the same measurement returns 21 of 371 near-domain and 27 of 218 legal segments; the legal figure reflects formulaic legal phrasing recurring across disjoint articles rather than content leakage, and it is the mirror image of the yrl side, where orthographic

variation suppresses exact recurrence.

What our speech numbers are computed over.

Two properties of the time-aligned speech source qualify the speech scores we report. First, its alignment unit is the pause-delimited *turn*, not the sentence (median 1.8 s and 4 words, a third of them two words or fewer; only 10% end in sentence-final punctuation), so most segments are mid-utterance fragments, a unit at once easier than a sentence (short) and harder (fragmentary, context-dependent). Second, its Portuguese side is a field free-translation written by the transcribers: regional Upper Rio Negro and/or L2 Portuguese (sentence-final discourse *nem* in 13.0% of the speech test set, non-standard spelling 9.6%, non-standard agreement 2.6%; the constitutional set scores 0% on all three). Into Portuguese, some model outputs use standard Portuguese forms where the references use regional or non-standard forms, which chrF++ can penalize sentence by sentence: “*Lili, aonde está*” (reference) vs. “*Lili, onde está*” (system) scores 64.3, and 100.0 against a standardized reference. Aggregated the effect vanishes (standardizing unambiguous spellings moves 12 runs by -0.61 to $+0.62$, mean ≈ -0.07), so no headline number needs correcting; we leave the references untouched, which is also the CARE-consistent choice: “correcting” Indigenous speakers’ Portuguese is not ours to do. In the pt→yrl direction the same text is *input*; robustness to standardized Portuguese input remains untested.

Ethics Statement

Our process is informed by the ethical framework of Mager et al. (2023) and by CARE principles. We contacted the Academia da Língua Nheengatu but have not obtained a response, endorsement, or institutional approval; silence is not treated as consent. The glossary and orthographic choices in this work are therefore provisional and source-derived rather than authoritative. Future governance should involve speakers from multiple regions and orthographic traditions. Licensing is layered: official-text layer outside copyright protection under Article 8 of Brazilian Law 9,610/1998; the treebank and speech layers inherit CC BY-NC-SA; the dictionary-example layer (Avila, 2021) is redistributed with the author’s written permission; the remaining third-party layers ship as reproducible extraction pipelines pending permission; the three NLLB-derived checkpoints are publicly available

on Hugging Face under CC-BY-NC-4.0; the Qwen-derived checkpoint is released under Apache-2.0. We position the system as writing/reading support, never official translation.

Speaker evaluation protocol. No speaker evaluation has been run for this version; the instrument below is built and being fielded, and its results will be reported separately. It presents 75 items in a self-contained offline form: absolute adequacy/naturalness on 30 system outputs balanced across legal, near-domain and speech registers, interleaved with **9 hidden human-reference anchors** (23% of the block, WMT-style) that give a calibration ceiling and an inattentive-rater check; 20 blind pairwise items (free vs. constrained decoding, with preference asked *before* the target term is revealed); 12 blind system-vs-official-translation items; and 4 open language questions. Item counts follow a power analysis. Speech items carry their neighbouring conversational turns as explicitly non-evaluated context, and evaluators are told the Portuguese of those turns is the transcribers’ own regional writing, not a translation error. Evaluators are told translations come from undisclosed mixed sources, never that all are machine-made, and are **debriefed within the form** about the hidden human references and why they were concealed; blinding keys never ship with the questionnaire. Evaluation data are anonymous by default. Any option to be acknowledged by name is collected through a separate, explicit opt-in that is not linked to the participant’s individual ratings. Participation is voluntary and compensated, with informed consent obtained before rating and the right to withdraw data until analysis. Ratings are stored under pseudonymous identifiers with restricted access. No speaker data have been analysed for this submission.

Acknowledgments

We thank Marcel Twardowsky Avila for permission to redistribute the dictionary example pairs derived from his thesis; the Revista Leetra Indígena at UFSCar and the authors of its bilingual school books; and the fifteen Indigenous translators of the Upper Rio Negro and Middle Tapajós whose translation of the Constitution, published by the STF and the CNJ, anchors this corpus.

References

- Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Doctoral thesis, estudos da tradução, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.
- Seth Aycok, David Stap, Di Wu, Christof Monz, and Khalil Sima’an. 2025. *Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?* *Preprint*, arXiv:2409.19151. ICLR 2025 (Spotlight).
- Toms Bergmanis and Mārcis Pinnis. 2021. *Facilitating terminology translation with target lemma annotations*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.
- Aline da Cruz. 2011. *Fonologia e Gramática do Nheengatú: A língua geral falada pelos povos Baré, Warekena e Baniwa*. Ph.D. thesis, Vrije Universiteit Amsterdam. LOT Dissertation Series.
- Wilmar da Rocha D’Angelis. 2023. *A língua Nheengatu e suas ortografias: questões técnicas e de política linguística*. *LIAMES: Línguas Indígenas Americanas*, 23:e023004.
- Leonel Figueiredo de Alencar. 2024. *A Universal Dependencies treebank for Nheengatu*. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 37–54, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. *Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas*. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.
- Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. *Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages*. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

- Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montañó, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. [Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages](#). In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.
- Javier Garcia Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. [BSC submission to the AmericasNLP 2024 shared task](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 143–149, Mexico City, Mexico. Association for Computational Linguistics.
- Isabel Gonçalves, Paulo Cavalin, and Claudio Pinhanez. 2025. [Exploring performance variations in finetuned translators of ultra-low resource languages: Do linguistic differences matter?](#) *Preprint*, arXiv:2511.22482.
- Edward Gow-Smith and Danae Sánchez Villegas. 2023. [Sheffield’s Submission to the AmericasNLP Shared Task on Machine Translation into Indigenous Languages](#). *Preprint*, arXiv:2306.09830. Also ACL Anthology 2023.americasnlp-1.21, pp. 192–199.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Jonathan Hus, Antonios Anastasopoulos, and Nathaniel Krasner. 2025. [Machine translation using grammar materials for LLM post-correction](#). In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. [Bilex Rx: Lexical Data Augmentation for Massively Multilingual Machine Translation](#). *Preprint*, arXiv:2303.15265.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024b. [Navigating the metrics maze: Reconciling score magnitudes and accuracies](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.
- Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. [Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.
- Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. [Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas](#). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.
- MihaiPopa-1. 2026. [OmniTranslate 1.0](#). Hugging Face model card. Accessed September 11, 2026.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, et al. 2022. [No Language Left Behind: Scaling Human-Centered Machine Translation](#). *Preprint*, arXiv:2207.04672.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, et al. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Omnilingual MT Team. 2026. [Omnilingual MT: Machine Translation for 1,600 Languages](#). *Preprint*, arXiv:2603.16309.
- Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 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, Vienna, Austria. Association for Computational Linguistics.
- Claudio Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. [Human evaluation of the usefulness of fine-tuned English translators for the Guarani Mbya and Nheengatu Indigenous languages](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 32–36, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Matt Post and David Vilar. 2018. [Fast lexically constrained decoding with dynamic beam allocation for neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.
- Antonio Carlos Quinto. 2025. [Projeto baseado em IA atua na preservação da língua indígena dheengatu, no Alto Rio Negro, no Amazonas](#). *Jornal da USP*. Speaker estimate attributed to Thomas Finbow, University of São Paulo.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#). Model card: <https://huggingface.co/Qwen/Qwen3.5-2B>.
- Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. [Findings of the WMT 2023 shared task on machine translation with terminologies](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.
- David Stap and Ali Araabi. 2023. [ChatGPT is not a good indigenous translator](#). In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.
- Supremo Tribunal Federal and Conselho Nacional de Justiça. 2023. [Mundu Sa Turusu Waá: Ubêuwa Mayé Míra Itá Uikú Arâma Purâga Iké Brazíu Upé](#). Conselho Nacional de Justiça, Brasília. Translation of the 1988 Constitution of the Federative Republic of Brazil into Nheengatu.
- 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](#). *Preprint*, arXiv:2309.16575. ICLR 2024.
- Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. [X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale](#). *Preprint*, arXiv:2410.03115. ICLR 2025 (Spotlight).
- Chen Zhang, Xiao Liu, Jiaheng Lin, and Yansong Feng. 2024a. [Teaching large language models an unseen language on the fly](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.
- Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. [Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

A Reproducibility

Hardware and budget. A 20k-step 600M run takes 72–77 minutes on a single NVIDIA H100 (≈ 1.3 GPU-h); the 1.3B flagship needs about four hours of wall clock. Jobs killed by a wall-clock limit resume from the last checkpoint, and we note one pitfall we hit: early stopping restarts its patience counter on resume, so patience is not comparable across interrupted and uninterrupted runs.

Training. Base checkpoints are NLLB-200-distilled 600M and 1.3B. The new language token `yrl_Latn` is initialized from the `grn_Latn` (Guarani) embedding, a starting point only; mixing Guarani training data was measured and rejected (-0.26 chrF++). Hyperparameters follow AmericasNLP submissions in the same line (Gow-Smith and Sánchez Villegas, 2023; García Gilabert et al., 2024): full fine-tuning (no adapters), $lr\ 1 \times 10^{-4}$ with inverse-square-root schedule and 400 warmup steps, weight decay 0.01, no label smoothing, max sequence length 192, effective batch 32 (16×2 accumulation), bf16, 20k steps, evaluation every 1000 steps, best checkpoint selected by dev chrF++ with early-stopping patience 8, beam 4 for the in-training generation.

The architecture control. The decoder-only side of §5.5 is Qwen/Qwen3.5-2B (Qwen Team, 2026), the instruction-tuned release rather than the pre-trained base, trained on the same v3 data, with the same three seeds, and scored with the same decoding and metrics as the seq2seq runs. The original Qwen checkpoint is a vision-language foundation model whose card counts 2B parameters in the language model; we load it with the text-only Qwen3_5ForCausalLM class, which instantiates the language backbone alone (1.882B parameters, the vision tower never constructed) and we use its own tokenizer (248,320 pieces, tied input and output embeddings). Adaptation is full fine-tuning with no adapters: $lr\ 2 \times 10^{-5}$ with cosine schedule and 400 warmup steps, weight decay 0.01, fused AdamW, fp32 master weights under bf16 autocast, gradient checkpointing, maximum sequence length 512 for prompt and target together, effective batch 32 (8×4 accumulation), 20k maximum steps, evaluation every 1000 steps on a 600-example development subset, best checkpoint by development loss (the seq2seq side keeps the checkpoint with the best development chrF++; Table 7 scores both selections). Training pairs are rendered with a plain Portuguese instruction template rather than the chat template (*Traduza do <source> para o <target>.*, then the source line and an empty target line), and the reasoning-delimiter tokens are suppressed at decoding so that generation budget cannot be spent on them. In all three seeds the best checkpoint is step 1,000 (1.15 epochs over the 27,900 bidirectional training examples) and the remaining 19k steps never improve development loss. Development chrF++ nonetheless keeps improving well

cell	seq2seq 600M	dec.-only @1k	dec.-only late	Δ late	3-SE
legal pt→yrl	29.57±0.22	25.89±0.67	27.67±0.57	+1.90	1.06
legal yrl→pt	34.83±0.29	28.14±0.68	33.11±1.51	+1.72	2.67
near-dom. pt→yrl	60.87±0.49	52.11±0.82	62.37±0.54	-1.50	1.26
near-dom. yrl→pt	56.68±0.23	48.94±0.51	59.91±0.05	-3.23	0.41
speech pt→yrl	46.23±0.48	39.96±1.11	44.74±0.90	+1.49	1.77
speech yrl→pt	50.95±1.05	47.16±1.88	47.88±1.60	+3.08	3.31

Table 7: The architecture control under both checkpoint selections (development sets, chrF++, mean±sd over three seeds). “@1k” is the loss-selected checkpoint; “late” is steps 11–12k, scored after submission as a proxy for chrF++ selection. Δ is seq2seq minus late decoder-only (positive favours the seq2seq); 3-SE is three standard errors of the difference, the pre-specified threshold for that comparison, and a cell is conclusive when $|\Delta| \geq 3\text{-SE}$ and $|\Delta| \geq 1.0$: legal pt→yrl (seq2seq) and both near-domain cells (decoder-only).

past that point (Table 7): at steps 11–12k the three seeds reach 27.67 ± 0.57 on legal pt→yrl against 25.89 ± 0.67 at step 1,000, and 62.37 ± 0.54 against 52.11 ± 0.82 on near-domain pt→yrl. The conclusiveness threshold is three standard errors of the difference between the two three-seed means, computed separately for each set, direction and metric; across the six cells of the loss-selected comparison it ranges from 0.98 to 3.73 chrF++. Speech yrl→pt clears it by 0.06 chrF++ but falls 0.19 short under spBLEU, which is why the cell is reported and not claimed.

Software. Metrics are computed with sacreBLEU 2.6.0 (nrefs:1, case:mixed; chrF++ with nc:6, nw:2, space:no; spBLEU with tok:flores200); NLLB models are trained with Transformers 4.55.4 and PyTorch 2.6.0, seeds 1–3. The Qwen control uses a separate environment with Transformers 5.14.1. Confidence intervals use paired bootstrap resampling with 1,000 resamples.

Evaluation. chrF++ (Popović, 2017) is sacreBLEU’s (Post, 2018) CHRF with word_order=2 and the default $\beta = 2$. Table 8 reports the full test results additionally in plain chrF (word_order=0) and spBLEU (tokenize=flores200), so that they can be read against work using either convention: much of the Indigenous-language literature reports plain chrF, and spBLEU is the severest of the three on a morphologically rich target. We deliberately use *no* neural metric: COMET, CometKiwi and MetricX have not been validated for yrl and provide no dedicated language coverage, so their scores here would be uninterpretable. TSR (§4) is scored over the union of the audited glossary and the base bilingual lexicon, with a yrl lemma table, a fuzzy

set	dir.	chrF++	chrF	spBLEU
<i>flagship</i>				
legal	pt→yrl	31.70	35.95	12.83
legal	yrl→pt	37.39	39.56	19.90
near-dom.	pt→yrl	64.11	65.99	54.65
near-dom.	yrl→pt	57.10	57.86	41.68
speech	pt→yrl	45.24	49.77	27.86
speech	yrl→pt	49.67	50.88	31.71
<i>avg</i>				
legal	pt→yrl	32.78	36.93	13.43
legal	yrl→pt	38.14	40.29	20.73
near-dom.	pt→yrl	64.54	66.44	54.85
near-dom.	yrl→pt	58.76	59.57	43.49
speech	pt→yrl	47.45	51.89	30.56
speech	yrl→pt	49.37	50.66	31.57

Table 8: Final test under three metrics. Plain chrF is approximately 0.8–4.5 points above chrF++ in these results; chrF++ additionally incorporates word n -grams. spBLEU scores are lower on these sets. The ordering of the two systems within each set and direction is the same under all three metrics; comparisons across directions need not have the same ordering.

threshold of 0.85 applied only to forms of five characters or more, and single-word Portuguese terms under four characters skipped. Each model–test combination was decoded once, under a written protocol fixing the system, the checkpoint and the metric before decoding. The legal and near-domain sets are reused holdouts, first evaluated in Phase 1 and once in Phase 2; `test_fala` was spent in the Phase-2 pass.

Audit categories and residual estimates. Two v3 sources were audited on samples of 200 and 240 pairs, with evidence attached to every judgement: 200 speech pairs, scored against the temporal context of neighbouring turns, and 240 dictionary-example pairs, scored against a window of the page’s source text.

	ok	minor	suspect	defective
Speech (200)	183	11	2	4
Dictionary ex. (240)	208	16	—	16

Table 9: Audit outcomes; the approval rates quoted in §3 are the *ok* column.

The residual estimates are lower than the non-ok mass in Table 9 because the two tails were treated differently. In the speech sample all four defective pairs were code-switch identity pairs, a mechanically detectable class that we then removed population-wide (42 pairs across the full resource); what survives is the suspect category, $2/200 = 1.0\%$

(Wilson 95% CI $[0.3, 3.6]$). In the dictionary sample the 31 corrections returned by the audit were verified against the page sources and applied to the audited pairs only, so the observed defective fraction is $16/240 = 6.7\%$ (Wilson interval $[4.1, 10.6]$). Because the sampling procedure was not preserved, these sample proportions and nominal intervals do not establish population error bounds for the unaudited pairs. Both residual classes are distinct from mispairing: 28 of the 32 dictionary problems are left-edge truncation or bleed on the yrl side, the Portuguese side is correct in 239/240, and cross-pairings are zero.

The French–yrl pairs. The 925 FR–yrl pairs are concatenated with the pt–yrl training file before the bidirectional expansion, so both directions are trained: 925 fra→yrl and 925 yrl→fra, 1,850 of the 27,900 training examples, and 12,467 yrl running words, 9.4% of the yrl side the model is fit to. They carry NLLB’s own fra_Latn and yrl_Latn tags and no additional source or register tag. There is no upsampling, downsampling or per-source balancing: the concatenated set is shuffled once, so every source contributes in proportion to its size. Inline terminology annotation is applied only to por_Latn sources, so the French pairs are never annotated. yrl→fra is trained and never evaluated; the mirroring is applied uniformly to every training file rather than decided per source.

Glossary provenance. The 904 training-glossary entries were mined from the training pairs and from external lexical resources. Evaluation references enter the pipeline only as an exclusion filter, deleting mined candidates that coincide with a development or test pair; they are never a source of entries. Checking every entry against the Portuguese side of both partitions, none has support only in the evaluation references.

Statistics. Beyond the ± 0.35 chrF++ ruler of §4, the TSR rulers on the legal dev set are $\pm 1.6\text{pp}$ unhinted and $\pm 2.7\text{pp}$ hinted (3 seeds, 600M): a single-run TSR difference below $\sim 3\text{pp}$ is not interpretable, and we mark such comparisons as inconclusive instead of reporting them as effects. Of the two single-seed results that failed to replicate, one was a TSR. Paired comparisons of degeneration rates use McNemar’s exact test. The metric values covered by the recorded audit rounds were recomputed from stored hypothesis files independently of the training-time logs. The v2 speech scores are

retained evaluation summaries; their hypotheses were not preserved. The released code includes the terminology matcher and training glossary. The base bilingual lexicon, the scoring lexicon’s other component, contains sentence-length entries from sources whose redistribution is still pending.

Code, data and models. Four checkpoints are released on Hugging Face: the NLLB 600M annotation-plus-token-pair model (seed 3), the 1.3B flagship and checkpoint average, and the text-only Qwen model (seed 2, step 11,000). The Qwen release has the highest near-domain pt→yrl development chrF++ among the three late checkpoints. Table 7 reports three-seed architecture-control results; its NLLB 600M comparator uses no terminology mechanisms. The final-test results in Table 5 concern the flagship and average. Model cards and weights are linked at <https://huggingface.co/voz-ancestral>. The extraction, alignment, annotation, constrained-decoding and evaluation scripts, and the corpus layers whose licences permit redistribution are at <https://github.com/fabriciocarraro/NheengatuMT>. Third-party layers ship as reproducible extraction pipelines instead of redistributed text where permission is pending; the dictionary-example layer is included with its author’s written permission.