

What the Macro Average Hides in Multilingual Translation Evaluation

Arman

Independent Researcher, Doha, Qatar

Abstract—Multilingual translation models are routinely reported as a single aggregate score across dozens or hundreds of languages, a practice that conceals two distinct problems. The first is well documented: low-resource languages receive lower quality. The second is less examined and arguably more consequential: low-resource languages may fail differently rather than merely worse, producing outputs in the wrong language entirely, corrupting numeric content, or degenerating into repetition, while high-resource languages fail gracefully through awkward phrasing. A translation that is clumsy remains usable; one that silently alters a dosage, a date, or a language does not. This report specifies EquiTranslate, a pre-registered protocol for separating quality inequality from reliability inequality within a single many-to-many model. We evaluate NLLB-200 (distilled, 600M) on 300 FLORES-200 sentences across eight languages spanning three resource tiers, four scripts, and four language families, in eleven translation directions including two non-English pairs. Quality is measured with chrF++, spBLEU under the shared FLORES-200 SentencePiece tokenizer, and COMET where its backbone supports the language. Reliability is measured by five fully automatic detectors applied to every sentence and validated against a blinded, stratified human annotation sample that measures detector precision and recall rather than assuming it. We additionally compare direct translation against routing through English for non-English pairs, and quantify the tokenization cost imposed on each script. The contribution is a protocol and a set of pre-registered hypotheses, not a new model. Its claim is methodological: a fairness evaluation reporting only macro-averaged quality cannot distinguish a system that is uniformly mediocre from one that is excellent in nine languages and unusable in the tenth.

Index Terms—multilingual machine translation; low-resource languages; translation evaluation; algorithmic fairness; hallucination detection; EquiTranslate

I. INTRODUCTION

A model advertised as supporting two hundred languages invites a question that a single number cannot answer: supporting them how well, and how well for whom? NLLB-200 translates between roughly two hundred languages [1], and FLORES-200 provides professionally translated, sentence-aligned evaluation data across that same set [1, 2]. Together they make cross-language comparison technically straightforward. What remains unresolved is what should be compared.

The standard practice is to report a macro average across languages, sometimes with a per-language appendix. This framing treats a language as a row in a table and quality as a scalar. It produces a well-known and repeatedly confirmed observation, namely that languages with less training data score lower, and stops there. Reproducing that observation is not a research contribution.

This protocol proceeds from a different question. Holding the model fixed, does the character of failure change with resource level? A translation can be wrong in ways that differ enormously in cost. Awkward word order in French is a cosmetic defect. A Sinhala output silently rendered in English, or a Yoruba output in which a quantity has changed, is a failure of a different kind: confidently presented, invisible to a monolingual user, and potentially harmful in a health, legal, or safety context. If such failures concentrate in exactly the languages whose speakers have the fewest alternative translation options, then the equity problem is sharper than the quality gap suggests, and it is invisible to chrF.

1.1 Contributions

- A pre-registered protocol separating quality inequality from reliability inequality, with hypotheses stated before data collection and explicit falsification criteria for each.
- Five language-agnostic automatic reliability detectors requiring no knowledge of the target language, applied at full coverage rather than to a hand-annotated subset, with a blinded validation design measuring their precision and recall instead of assuming it.
- A measurement of direct versus English-pivot routing for non-English language pairs, in which the pivot hypothesis is tested rather than assumed.
- A tokenization-cost analysis framing subword inflation as an access and pricing problem,

measured as tokens per sentence relative to English for identical semantic content.

- Released code, fixed configuration, captured environment metadata, and pre-specified tables, permitting the protocol to be re-run against other models without redesign.

1.2 Scope and Non-Goals

This study does not train a translation model, does not propose an architecture, and does not claim state-of-the-art quality. It evaluates one publicly available checkpoint under a fixed protocol. Interventions requiring pretraining-scale compute are explicitly out of scope; Section 5.2 records which elements of the originating project proposal were removed for this reason and why.

1.3 Research Questions

Table 1: Research questions and the sections addressing them.

	Question	Addressed by
RQ1	How large is the translation-quality gap between resource tiers when measured with tokenizer-neutral metrics?	§3.5, H1
RQ2	Does the character of failure, not only its frequency, differ by resource tier?	§3.6, H2, H3
RQ3	For non-English language pairs, does routing through English help or harm?	§3.4.3, H4
RQ4	What tokenization cost does the shared vocabulary impose on each script for identical content?	§3.5.4, H5
RQ5	Can language-agnostic automatic detectors substitute for human annotation at measurable precision and recall?	§3.7, H6

II. LITERATURE REVIEW

2.1 Massively Multilingual Translation and Its Benchmarks

NLLB-200 was released as a many-to-many model covering roughly two hundred languages, accompanied by FLORES-200, an evaluation set of professionally translated sentences aligned across all supported languages [1]. Its predecessor established the sentence-aligned, many-language evaluation format on which this work depends [2]. MADLAD-400 extends coverage further, auditing a document-level corpus spanning 419 languages and training translation models over more than 450 [5]. WMT24++ expands human-written references and post-edits to 55

languages and dialects, and offers a useful secondary check on whether findings obtained on FLORES generalise beyond it [4].

Sentence alignment across languages is the property making this study possible: because every language in FLORES contains a translation of the same source sentence, differences in score cannot be attributed to differences in test content. This protocol preserves that property strictly, evaluating every direction on an identical set of sentence identifiers.

2.2 Hallucination and Omission

HalOmi provides manually annotated hallucination and omission labels across eighteen translation directions with varying resource levels and scripts, and

demonstrates that conclusions drawn from a single language pair do not hold at multilingual scale [3]. That finding motivates the present design directly: a reliability result obtained on English–German says little about English–Sinhala, and a protocol validating detectors on one pair and deploying them on ten is unsound.

This work differs from HalOmi in method rather than aim. HalOmi contributes human labels; this protocol contributes automatic detectors at full coverage plus a blinded human sample sized to validate them. The two are complementary, and HalOmi's definitions of hallucination, being added content unsupported by the source, and omission, being source content absent from the output, are adopted here without modification.

2.3 Interference and Negative Transfer

Shared multilingual parameters permit transfer between related languages but also allow dominant languages to consume model capacity, degrading performance on others [16]. Temperature-based sampling of the training distribution is the standard mitigation [12], and language-specific low-rank adapters offer a parameter-efficient alternative [13]. Both operate at training time. This protocol measures the consequences of such choices as they appear in a

released checkpoint; it does not attempt to intervene in them, for reasons set out in Section 5.2.

2.4 Tokenization as a Cost and Access Problem

Subword vocabularies trained on skewed multilingual corpora segment under-represented scripts into more tokens for identical content. This inflation has been shown to translate directly into higher API cost, reduced effective context, and slower inference for speakers of those languages [20, 21]. The measurement is simple and the implication is concrete, which is why it is included here: a quality gap is contestable, but a speaker paying two or three times more for the same sentence is an arithmetic fact.

III. METHODOLOGY

3.1 Language Sample

Eight languages were selected to vary resource tier, script, and family while remaining small enough that every direction can be run to completion within a single free-tier GPU session. The originating proposal suggested twelve to twenty languages; eight is chosen deliberately, because a completed eight-language study with confidence intervals is worth more than a twenty-language study abandoned at the annotation stage.

Table 2: Language sample. Script and family are varied within tier so that a tier effect cannot be confounded with a single script or family.

Language	FLORES code	Script	Tier	Family	Role in design
English	eng_Latn	Latin	High	Indo-European	Pivot and reference point
French	fra_Latn	Latin	High	Indo-European	High-resource ceiling
Japanese	jpn_Jpan	Japanese	High	Japonic	High resource, non-Latin script
Hindi	hin_Deva	Devanagari	Medium	Indo-European	South Asian comparison
Bengali	ben_Beng	Bengali	Medium	Indo-European	Related-pair transfer with Hindi
Swahili	swh_Latn	Latin	Medium	Niger-Congo	African, Latin script
Sinhala	sin_Sinh	Sinhala	Low	Indo-European	Low resource, unique script
Yoruba	yor_Latn	Latin	Low	Niger-Congo	Low resource, diacritics and tone

Two deliberate controls are embedded in this set. Japanese is high-resource but non-Latin, separating the effect of script from the effect of resource level.

Yoruba and Swahili share both script and family but differ in tier, separating resource level from typology.

3.2 Translation Directions

Eleven directions are evaluated: six out of English, three into English, and two between non-English

languages. The last group is included because it is where pivot routing becomes a live question and where interference effects are least studied.

Table 3: Translation directions. The two non-English pairs are additionally run through an English pivot, giving thirteen evaluated system-direction combinations.

Direction	Group	Purpose
eng → fra	Out of English	High-resource ceiling
eng → hin	Out of English	Medium resource, Devanagari
eng → ben	Out of English	Medium resource, Bengali script
eng → swl	Out of English	Medium/low resource, Latin
eng → sin	Out of English	Low resource, unique script
eng → yor	Out of English	Low resource, diacritics
hin → eng	Into English	Reverse direction asymmetry
swl → eng	Into English	Reverse direction asymmetry
sin → eng	Into English	Reverse direction asymmetry
hin → swl	Non-English	Direct versus pivot
swl → hin	Non-English	Direct versus pivot

3.3 Data

FLORES-200 is the sole evaluation source. The dev split is used for all development, debugging, and threshold setting. The devtest split is used exactly once, to produce the final reported numbers. This separation is the single most important procedural commitment in this document: repeated evaluation against devtest converts it into a training set by selection, and any improvement reported afterwards is

an artifact of that selection rather than a property of the system.

Each direction is evaluated on the same 300 sentences, identified by FLORES sentence identifier rather than by row position, and only identifiers present in every one of the eight languages are retained. Equal sample sizes across directions are a precondition for the macro average and the tier gap in Section 3.5.3 to be interpretable.

Table 4: Data configuration.

Parameter	Value	Rationale
Evaluation set	FLORES-200 devtest	Professionally translated, sentence-aligned across all languages
Development set	FLORES-200 dev	All tuning and threshold setting; devtest untouched until final run
Sentences per direction	300 (identical ids)	Equal n required for macro averaging; fits free-tier GPU budget
Selection rule	First 300 aligned ids	Deterministic and citable; no random seed dependence
Fine-tuning data	None	No training is performed; see Section 5.2

Parameter	Value	Rationale
License	FLORES-200: CC-BY-SA 4.0 NLLB checkpoints: CC-BY-NC 4.0	Non-commercial restriction applies to any redistributed model artifact

3.4 Systems Under Comparison

Three systems are evaluated. All use the same checkpoint and differ in decoding or routing only. This is a deliberate narrowing of the originating five-system design, discussed in Section 5.2.

Table 5: Systems under comparison. System C is designated optional-if-time-permits; Systems A and B constitute the core study and are sufficient to address RQ1 through RQ5.

System	Description	Differs in	Directions
A — Baseline	NLLB-200 distilled 600M, beam search, no modification	—	All 11
B — English pivot	Two-hop translation L1 → English → L2, same checkpoint	Routing	2 non-English pairs
C — Constrained decoding	Numbers and named entities constrained during generation	Decoding	All 11

3.4.1 System A: Baseline

NLLB-200 distilled 600M, beam search with four beams, greedy tie-breaking, maximum 256 input and 256 generated tokens, float16 on GPU. The target language is forced through the decoder start token. This resolution is version-sensitive: the tokenizer attribute used in most published example code was removed in a 2024 release of the Transformers library, and code resolving the language code through the general vocabulary lookup will silently return the unknown-token identifier, causing the model to translate into an arbitrary language. The implementation raises an explicit error in this case rather than producing a score. This is recorded because a silent version-dependent failure of this kind would be indistinguishable in the results from a genuine low-resource quality collapse.

3.4.2 System B: English Pivot

For the two non-English pairs, the source is translated into English and the English output is then translated into the target. The intermediate English text is retained for every sentence, because attributing a failure to the first or the second hop is impossible without it.

3.4.3 The Pivot Hypothesis

English-centric training makes English-involving directions the best-supported paths through a many-to-many model, which is the usual argument that pivoting should help non-English pairs. The counter-argument is that a pivot introduces a second lossy step and forces the output through English lexical and grammatical categories, so that information absent from English, such as grammatical gender, evidentiality, honorific level, or noun class, is discarded at the first hop and cannot be recovered at the second. Both arguments are plausible. Neither is measured in the originating proposal, which assumes the first. This protocol measures it.

3.5 Quality Metrics

chrF++, being the character n-gram F-score with word bigrams at word_order = 2, is the primary quality metric [7, 8]. Character n-grams degrade gracefully across morphologically rich, agglutinative, and unsegmented languages, where word-level overlap approaches noise. It is computed identically for every language in the study.

Table 6: Secondary metrics and their configuration.

Metric	Configuration	Note
spBLEU	sacreBLEU with the flores200 SentencePiece tokenizer [6]	The default 13a tokenizer is built for European scripts and inflates the apparent gap for Sinhala and Japanese by measuring segmentation rather than translation
BLEU (Japanese)	ja-mecab tokenizer	Applied only where Japanese is scored in isolation
COMET	Unbabel/wmt22-comet-da [9, 10]	Reported only where the XLM-R backbone [11] covers the target language
Efficiency	Tokens per sentence; seconds per sentence; output-to-reference length ratio	Reported per direction alongside quality

COMET coverage. The wmt22-comet-da model is built on XLM-R large, whose pretraining covers 100 languages [11]. Yoruba is not among them. COMET will nonetheless emit a well-formed score for Yoruba, and that score is not comparable to the others. It is therefore reported as unavailable with an explicit note rather than as a number, and Yoruba is compared on chrF++ and spBLEU only. Reporting a metric outside its validated domain is the most direct way to invalidate a fairness conclusion, because the language

most likely to be mis-scored is the one the conclusion concerns.

3.5.3 Fairness Aggregates

For L evaluated languages with scores S_1 through S_L , the following are reported for every system. The macro average is reported last and weighted least, because it is the statistic this study argues against relying upon.

Table 7: Fairness aggregates. An intervention raising the macro average while leaving the worst-language score unchanged has not achieved the stated goal, and the reporting order reflects this.

Statistic	Definition	What it detects
Worst-language score	$S_{\text{worst}} = \min(S_1 \dots S_L)$	Whether any language is left unusable
Tier gap	$\text{mean}(S \mid \text{high tier}) - \text{mean}(S \mid \text{low tier})$	Systematic stratification by resource level
Inequality	$SD(S_1 \dots S_L)$	Dispersion independent of the mean
Macro average	$(1/L) \sum S_i$	Overall level; reported for comparability, not as headline

3.5.4 Tokenization Cost

Because FLORES sentences are semantically parallel, the mean number of subword tokens per sentence in each language is a direct measure of the cost the shared vocabulary imposes for identical content. Inflation is reported as a ratio relative to English. The unit matters for interpretation: a ratio of 2.5 means a speaker of that language pays 2.5 times as much per sentence under token-based pricing, fits 40 per cent as much into a fixed context window, and waits proportionally longer for generation.

3.6 Reliability Detectors

Five detectors run over every translated sentence. All are language-agnostic, deterministic, and require no knowledge of the target language, which permits full coverage rather than sampling. They are proxies for the underlying phenomena, not ground truth, and Section 3.7 specifies how their precision and recall are measured rather than assumed.

Table 8: Automatic reliability detectors. Thresholds are set on the dev split and frozen before devtest is touched.

Detector	Definition	Detects	Known limitation
Off-target	Fewer than 50% of alphabetic characters belong to the target script (Unicode category)	Output produced in the wrong language entirely	Blind to Latin-to-Latin confusions; those fall to the copy detector
Source copy	At least 80% of output tokens also appear in the source	Model echoed the input instead of translating	Can fire on legitimately short or entity-heavy sentences
Numeric error	Multiset of numerals differs from the reference after normalising Devanagari, Bengali and fullwidth digits to ASCII	Quantities, dates and measurements altered	Cannot detect spelled-out numerals
Repetition	Any 4-gram occurs more than three times	Degenerate looping under beam search	May fire on legitimately repetitive source text
Length anomaly	Output-to-reference character ratio below 0.6 or above 1.6	Candidate omission or over-generation	Weakest detector; ratios vary legitimately by typology

Rationale for off-target detection. Producing the wrong language is the failure mode hand annotation systematically under-reports, because an annotator scoring adequacy on a five-point scale records it as a low score rather than a categorical failure, and an annotator who does not read the target language cannot flag it at all. It is also the failure with the clearest downstream harm: a monolingual user receives fluent text in a language they cannot read, with no signal that anything went wrong. It is detectable from Unicode script composition alone.

Rationale for reporting length ratio as a suspicion. A character ratio of 0.5 may be entirely correct when translating from an analytic language into an agglutinative one. This detector is used to rank sentences for human review, and its rates are reported as omission-suspect rather than as measured omission. The distinction is not cosmetic: reporting a proxy under the name of the phenomenon it proxies is how a methodological shortcut becomes a false empirical claim.

3.7 Human Validation Protocol

The originating proposal called for 50 to 200 hand-annotated sentences per language across the full language set, plus a second annotator per language for agreement statistics. At eight languages this is

between 800 and 3,200 judgements, in languages the investigator does not read, requiring sixteen qualified bilingual annotators. That design does not fail gracefully: it fails by being abandoned partway, leaving annotation coverage correlated with annotator availability, which itself correlates with resource level, producing a bias in exactly the direction the study seeks to measure.

The protocol is therefore inverted. Automatic detectors provide full coverage and constitute the reported result. Human annotation is used to measure how far those detectors can be trusted.

- Per direction, 25 detector-flagged and 25 unflagged sentences are drawn, giving 50 per direction and 550 in total.
- The two strata are pooled and shuffled before presentation, so the annotator cannot tell which sentences were flagged.
- Blinding is the point of the design. An unblinded annotator reviewing flagged sentences will confirm the detectors, and the resulting precision estimate is worthless.
- Where two annotators are available for a language, Cohen's kappa is reported [17]. Where only one is available, this is stated per language rather than silently omitted.

Table 9: Human annotation schema. Hallucination and omission follow HalOmi's definitions without modification [3].

Field	Scale	Definition
Meaning preservation	1–5	5 = fully accurate; 3 = understandable with noticeable errors; 1 = unusable or unrelated
Fluency	1–5	5 = natural and grammatical; 1 = unacceptable
Hallucination	0 / 1	Content present in the output that is unsupported by the source [3]
Omission	0 / 1	Content present in the source that is absent from the output [3]
Entity error	0 / 1	A named entity is corrupted, dropped, or substituted
Numeric error	0 / 1	A number, date, or measurement is altered
Safety-critical	0 / 1	The error would mislead in a health, legal, or safety context

For each detector, precision and recall against the human labels are reported, with the number of annotated items and the languages covered. A detector with high recall and moderate precision remains useful as a screening instrument, provided the paper says so. The failure to avoid is reporting detector rates as measured error rates without this validation step,

which would present an unvalidated heuristic as an empirical finding.

3.8 Statistical Analysis

All comparisons are pre-specified here. No test not listed below will be reported, and no threshold will be adjusted after devtest is unsealed.

Table 10: Pre-specified analysis plan.

Analysis	Method	Applied to
Uncertainty on every score	Paired bootstrap resampling, 1,000 resamples, 95% percentile interval [14]	Every segment-level metric in every results table
System comparison	Paired bootstrap on score differences over the same sentence ids	A versus B, A versus C within each direction
Categorical detector comparison	McNemar's test on paired binary outcomes	Flag rates between systems
Tier effect	Score $\sim$ System + Tier + Direction + System $\times$ Tier (mixed effects, language as random effect)	Whether an intervention helps tiers differently
Data volume versus quality	Spearman correlation, log-scaled x-axis	Reported training sentences against chrF++
Tokenization versus reliability	Spearman correlation	Token inflation against combined flag rate
Detector validation	Precision and recall against blinded human labels; Cohen's kappa where two annotators exist [17]	Section 3.7

On effect sizes. At 300 sentences per direction, a difference of one to two chrF++ points between systems generally falls within the bootstrap interval. Every reported difference is accompanied by its

interval, and any difference whose interval spans zero is described as inconclusive rather than as a small effect. This rule is fixed in advance because the

temptation to describe an inconclusive result as a trend arises only after seeing it.

3.9 Pre-Registered Hypotheses

The following are predictions, not results. Each is stated with the observation that would falsify it. The distinction is load-bearing: a protocol whose hypotheses cannot fail is not testing anything.

Table 11: Pre-registered hypotheses with falsification criteria.

	Hypothesis	Falsified if
H1	chrF++ decreases monotonically across high, medium and low tiers, with the high-to-low gap exceeding 15 points	The gap is under 15 points, or tier ordering is not monotonic under tokenizer-neutral metrics
H2	Off-target output rate is substantially higher for low-tier targets than high-tier targets	Off-target rates are comparable across tiers, or the tier ordering reverses
H3	Numeric corruption rate is higher for low-tier targets, and the tier gap in reliability flags is proportionally larger than the tier gap in chrF++	Numeric errors are tier-independent, or the reliability gap is no larger proportionally than the quality gap
H4	For non-English pairs, English pivoting raises fluency-weighted quality but produces at least as many omissions and entity errors as direct translation	Pivoting improves both quality and reliability, or degrades both
H5	Token inflation relative to English exceeds $2\times$ for at least one non-Latin script, and correlates positively with combined flag rate	All scripts fall under $2\times$, or the correlation is null or negative
H6	The off-target and numeric detectors achieve precision above 0.80 against blinded human labels	Either detector falls below 0.80 precision, in which case its rates are reported as screening output only

Note on H1. H1 is the least interesting hypothesis in the table and is expected to hold. It is included as a manipulation check: if the tier ordering does not reproduce, something is wrong with the pipeline, most plausibly a mis-resolved target language token as described in Section 3.4.1, and no other result should be interpreted until that is resolved. The contribution rests on H2 through H5.

IV. RESULTS

No measurements are reported in Section 4.2. The structure, column set, and row set of every table there are fixed before data collection so that the reported analysis cannot be reshaped after the fact to suit the outcome. Every cell marked with a dash will be filled from a single devtest run. Section 4.1 reports measured

values from prior published work, clearly attributed and not produced by this study.

4.1 Prior Published Evidence and Expected Effect Sizes

The table below contains measured values from published work, not from this study. It serves two purposes: it establishes that the effect sizes this protocol is powered to detect are plausible, and it provides a reference point against which the results of Section 4.2 can be compared once collected. Divergence between these figures and our own would itself be informative, since our checkpoint is the distilled 600M variant rather than the full-scale models most published numbers describe.

Table 12: Published reference values. All figures are cited from prior work; none were measured in this study. NLLB averages are over FLORES-101 for comparability with the cited source and are not directly comparable to per-direction FLORES-200 scores.

Quantity	Published value	Source and scope
NLLB-200 average, English $\rightarrow$ xx	34.0 spBLEU / 50.6 chrF++	NLLB-200 on FLORES-101 devtest, 101 languages, full-scale model [1]
NLLB-200 average, xx $\rightarrow$ English	41.2 spBLEU / 60.9 chrF++	As above; into-English directions score markedly higher [1]
NLLB-200 average, xx $\rightarrow$ yy	23.7 spBLEU / 41.4 chrF++	As above. The ≈ 19 chrF++ drop from into-English to non-English pairs frames the pivot experiment (H4) [1]
NLLB-200 gain over prior SOTA	+7.3 spBLEU ($\approx 44\%$ relative)	FLORES-101, 101 languages [1]
Tokenization length disparity	Up to $15\times$ between languages for identical content	Petrov et al., measured on FLORES-200 parallel sentences across commercial and multilingual tokenizers [21]
Byte- and character-level disparity	Over $4\times$ for some language pairs	Same study; the disparity is not an artifact of subword vocabularies alone [21]
Hallucination annotation coverage	18 directions, $\approx 2,865$ annotated examples	HalOmi; includes Yoruba, Kashmiri and Manipuri as low-resource cases [3]
Cross-pair generalisation of detectors	Single-pair conclusions do not hold at multilingual scale	HalOmi's central methodological finding; motivates per-language reporting in Section 3.7 [3]

Power implication. The ≈ 19 chrF++ spread between into-English and non-English directions reported in [1] is an order of magnitude larger than the one-to-two point resolution of a 300-sentence bootstrap interval. The study is therefore adequately powered for H1 and

H4. It is not powered to resolve small differences between decoding variants, which is why System C is designated optional and why Section 3.8 commits in advance to calling interval-spanning differences inconclusive.

4.2 Tables to be Filled from the Devtest Run

Table 13: Dataset and tokenization profile. Because FLORES sentences are semantically parallel, the final column is a direct measure of tokenization cost for identical content.

Language	Script	Tier	Sentences	Mean chars	Mean tokens	Inflation
English	Latin	High	300	—	—	1.00
French	Latin	High	300	—	—	—
Japanese	Japanese	High	300	—	—	—
Hindi	Devanagari	Medium	300	—	—	—
Bengali	Bengali	Medium	300	—	—	—
Swahili	Latin	Medium	300	—	—	—

Language	Script	Tier	Sentences	Mean chars	Mean tokens	Inflation
Sinhala	Sinhala	Low	300	—	—	—
Yoruba	Latin	Low	300	—	—	—

Table 14: Main translation results, System A. The Yoruba COMET cell is pre-marked unavailable rather than left to be filled, per Section 3.5.

Direction	Tier	chrF++ [95% CI]	spBLEU	COMET [95% CI]	s/sentence
eng → fra	High	—	—	—	—
eng → hin	Medium	—	—	—	—
eng → ben	Medium	—	—	—	—
eng → swl	Medium	—	—	—	—
eng → sin	Low	—	—	—	—
eng → yor	Low	—	—	n/a — coverage	—
hin → eng	Medium	—	—	—	—
swl → eng	Medium	—	—	—	—
sin → eng	Low	—	—	—	—
hin → swl	Cross	—	—	—	—
swl → hin	Cross	—	—	—	—

Table 15: Fairness aggregates by system. Column order is deliberate: worst-language performance first, macro average last.

System	Worst chrF++	Worst language	Tier gap	SD	Macro chrF++
A — Baseline	—	—	—	—	—
B — English pivot	—	—	—	—	—
C — Constrained decoding	—	—	—	—	—

Table 16: Reliability flag rates, percentage of sentences. This table, not Table 14, carries the central claim of the study.

Direction	Tier	Off-target	Src copy	Numeric	Repetition	Omis-susp.
eng → fra	High	—	—	—	—	—
eng → hin	Medium	—	—	—	—	—
eng → ben	Medium	—	—	—	—	—
eng → swl	Medium	—	—	—	—	—
eng → sin	Low	—	—	—	—	—
eng → yor	Low	—	—	—	—	—

Direction	Tier	Off-target	Src copy	Numeric	Repetition	Omis-susp.
hin → swh	Cross	—	—	—	—	—
swh → hin	Cross	—	—	—	—	—

Table 17: Direct versus English-pivot routing. Reporting quality and reliability side by side is required to test H4, which predicts they move in opposite directions.

Pair	chrF++ direct	chrF++ pivot	Δ [95% CI]	Omis. direct	Omis. pivot
hin → swh	—	—	—	—	—
swh → hin	—	—	—	—	—

Table 18: Detector validation against blinded human annotation. Tests H6.

Detector	Human label	Precision	Recall	n annotated	Languages
Off-target	Wrong language	—	—	—	—
Numeric error	Numeric error	—	—	—	—
Repetition	Hallucination	—	—	—	—
Length anomaly	Omission	—	—	—	—

Table 19: Qualitative error examples, selected after quantitative analysis to illustrate rather than to establish. All examples are drawn from FLORES-200 (CC-BY-SA 4.0), not from private or user-supplied text.

Language	Source	Reference	Model output	Error type	Severity
Sinhala	—	—	—	—	—
Yoruba	—	—	—	—	—
Swahili	—	—	—	—	—
Hindi	—	—	—	—	—

4.3 Planned Figures

Six figures are planned. Languages are ordered by resource tier throughout rather than alphabetically, so that the gradient is visible without reconstruction. Bootstrap intervals appear on every bar for which segment-level scores exist. All figures are rendered in

grayscale; where a second encoding channel is required, tier or system is distinguished by hatching rather than colour. No figure uses a radar chart, which compresses incommensurable axes onto a shared scale and invites visual comparisons the underlying numbers do not support.

Table 20: Planned figures.

	Figure	Type	Supports
1	Protocol pipeline: data → translation → metrics → detectors → validation → analysis	Schematic	Reproducibility
2	Mean subword tokens per sentence by language, annotated with inflation ratio	Grouped bars	H5, RQ4

	Figure	Type	Supports
3	chrF++ by target language and system, with bootstrap intervals	Grouped bars	H1, RQ1
4	Macro average, worst-language score and tier gap side by side per system	Three panels	RQ1, §3.5.3
5	Detector flag rates by language, stacked by detector	Stacked bars	H2, H3, RQ2
6	Direct versus pivot: quality and omission-suspect rate paired	Paired bars	H4, RQ3

Note on Figure 4. The three panels are presented at equal size deliberately. A reader consulting only the macro-average panel will draw a different conclusion from one consulting the worst-language panel, and the argument of this paper is that the second reader is correct.

V. DISCUSSION

5.1 Anticipated Contribution

If the pre-registered hypotheses hold, this study will establish that resource level predicts not only how well a multilingual translation model performs but how it fails, and that the second gradient is steeper than the first. That distinction changes what a responsible evaluation must report. A macro average is compatible with a system that is excellent in seven languages and produces confident, fluent, wrong-language output in the eighth, and the speakers of the eighth language are precisely those least able to detect it and least likely to have an alternative.

If the hypotheses fail, whether because reliability flags prove tier-independent or because the detectors prove unreliable against human labels, that is also a reportable result, and the protocol is written so that it can be reported without redesign. The claim being defended is not that low-resource translation is poor. It is that the field's default reporting practice cannot distinguish a system that is uniformly mediocre from one that is selectively unusable, and that these should not be scored alike.

5.2 Deviations from the Originating Proposal

The originating project proposal specified five experimental systems, twelve to twenty languages, a purpose-built benchmark, and human annotation at the scale discussed in Section 3.7. Several of those elements cannot be executed as specified. They are recorded here with reasons, because an undocumented silent reduction in scope is itself a reproducibility failure.

Table 21: Scope changes from the originating proposal, with reasons.

Removed or changed	Reason	Replacement
Retrieval-augmented translation	NLLB-200 is an encoder-decoder translation model with no instruction interface. Retrieved examples prepended to the input are translated as content, not treated as demonstrations. Retrieval-augmented MT with this architecture requires fuzzy-match fine-tuning, a separate project	Removed. If required, run on an instruction-tuned multilingual LLM and report it explicitly as a different model class
Balanced-data sampling as a standalone system	Temperature sampling is a pretraining data-mixture parameter [12]. It does not exist as a control on a released checkpoint. The proposal lists natural and balanced sampling as two	Removed as a standalone system; meaningful only as a LoRA fine-tuning variant, which is out of scope

Removed or changed	Reason	Replacement
	variants of the same frozen model, which cannot be constructed	
LoRA language adapters	Requires fine-tuning data, training compute, and a held-out tuning protocol beyond the budget of this study	Deferred; the protocol permits adapters to be added later as System D without redesign
GlobalMT-Fair benchmark construction	Building $\approx 1,500$ sentence meanings across 15 languages with entity, numeric, idiom and adversarial subsets requires professional translators in every language. It is a multi-year resource project, not a component of an evaluation study	Removed. FLORES-200 is used as-is, with its limitations stated in Section 5.3
Language sample of 12–20	Annotation and compute budget; a completed 8-language study with confidence intervals is stronger than an abandoned 20-language one	8 languages, chosen to vary script and family within tier
Default BLEU across languages	The 13a tokenizer is built for European scripts and would measure segmentation rather than translation quality for Sinhala and Japanese, inflating the apparent tier gap	spBLEU with the shared flores200 tokenizer; ja-mecab where Japanese is scored alone
COMET for all languages	The XLM-R backbone of wmt22-comet-da does not cover Yoruba [11]	COMET suppressed with an explicit note for Yoruba; chrF++ and spBLEU retained
Full human annotation at 50–200 per language	800–3,200 judgements across eight languages requiring sixteen bilingual annotators; the realistic outcome is partial coverage correlated with resource level, biasing the study in the direction it is measuring	Automatic detectors at full coverage, validated against a blinded 550-item stratified sample (Section 3.7)

5.3 Threats to Validity

Benchmark domain. FLORES-200 is drawn from web articles and is formal, edited, written prose. Conclusions drawn from it do not transfer automatically to conversational, code-switched, dialectal, or domain-specific text, and low-resource languages are plausibly over-represented in exactly the informal registers FLORES excludes. The tier gap reported here should be read as a lower bound on the gap in realistic use. Partial mitigation would be a secondary evaluation on WMT24++ [4], which covers literary, news, social and speech domains; this is listed as future work rather than claimed.

Reference quality. Low-resource reference translations in any benchmark are produced by a smaller pool of translators with less redundancy. If low-resource references are systematically noisier, a

portion of the measured quality gap is a property of the benchmark rather than the model. This protocol cannot separate the two, and does not claim to.

Metric validity under distribution shift. Neural metrics are trained predominantly on high-resource human judgements. Their reliability on low-resource output is itself a function of resource level, meaning the measurement instrument may be biased in the same direction as the phenomenon. This is why chrF++ rather than COMET is designated primary, and why the character-level and neural metrics are reported side by side: divergence between them is informative in itself.

Single model, single checkpoint. All results describe one distilled 600M checkpoint. Distillation may itself affect low-resource directions disproportionately. Nothing here generalises to larger NLLB variants, to

MADLAD-400 [5], or to LLM-based translation without re-running the protocol, which it is designed to permit.

Detector construct validity. The detectors measure surface properties correlating with the phenomena of interest. The off-target and numeric detectors have strong face validity; the length-ratio detector has weak face validity and is reported as a suspicion rate throughout. Section 3.7 quantifies rather than assumes this, and H6 commits in advance to demoting any detector falling below 0.80 precision to screening-only status.

Annotator availability. Qualified bilingual annotators will not be available for all eight languages. Coverage is reported per language rather than aggregated, and no agreement statistic is reported for a language annotated by one person.

5.4 Reproducibility and Ethics

- Every run records GPU model and memory, Python, PyTorch, Transformers and datasets versions, model and tokenizer revision, dataset split and sentence identifiers, decoding parameters, random seed, and git commit, written to a configuration file stored alongside the results.
- Code, configurations, per-sentence outputs, detector labels, and the annotation sheet are released. Translations are released with their corresponding FLORES identifiers so that any figure can be recomputed from source.
- Failed and unavailable directions are reported. A direction that could not be run is a result, not an omission.
- FLORES-200 is CC-BY-SA 4.0. NLLB-200 checkpoints are CC-BY-NC 4.0, which is non-commercial; any redistributed derivative inherits that restriction. Both license files are to be re-checked at publication time.
- No private, user-supplied, or personally identifying text is used. All qualitative examples come from the public benchmark.
- Annotators are credited by agreement and compensated; annotation guidelines are released with the data.

A note on framing. Results of this kind can be read as an indictment of a specific model. That reading would be unearned. NLLB-200 was released precisely to extend coverage to languages that had none, and measuring where it falls short is possible only because

it exists and is open. The finding this protocol anticipates concerns evaluation practice, namely that aggregate reporting hides reliability failures, and applies to any multilingual system evaluated the same way.

VI. CONCLUSION

This report specifies a pre-registered protocol for evaluating multilingual machine translation as a fairness-sensitive system, and commits in advance to the design, metrics, hypotheses, and analysis by which that evaluation will be judged. Its central methodological argument is that translation quality and translation reliability are separate quantities, that both stratify by resource level, and that the second may stratify more steeply than the first while remaining entirely invisible to the aggregate scores by which multilingual systems are conventionally reported.

The protocol is deliberately modest in scope: one checkpoint, eight languages, eleven directions, three systems, 300 sentences each. That modesty is a design choice rather than a limitation of ambition. A completed study whose every number carries a confidence interval, whose detectors are validated against blinded human judgement, and whose hypotheses were fixed before the data was seen, constitutes stronger evidence than a broader study assembled after the fact. The tables of Section 4.2 remain empty for the same reason: their structure is the commitment, and filling them is the experiment.

Once collected, the results will either support the claim that low-resource languages fail differently rather than merely worse, or they will not. Both outcomes are publishable under this design, and the protocol is written so that either can be reported without revision to anything preceding Section 4.

ACKNOWLEDGMENTS

This work was conducted independently, without institutional funding or compute support. The author thanks the bilingual annotators who contributed judgements under the protocol of Section 3.7; they are credited by name and language, where consent was given, in the released annotation guidelines. Any reviewers of this pre-registration are acknowledged here in the final version.

REFERENCES

Entries were checked against the ACL Anthology, arXiv, and NeurIPS proceedings. Verify page numbers and DOIs against the target venue's style guide before submission.

- [1] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, et al. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672, 2022.
- [2] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.
- [3] D. Dale, E. Voita, J. Lam, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, L. Barrault, and M. R. Costa-jussà. HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation. In *Proceedings of EMNLP 2023*, pages 638–653.
- [4] D. Deutsch, E. Briakou, I. Caswell, M. Finkelstein, R. Galor, J. Juraska, G. Kovacs, A. Lui, R. Rei, J. Riesa, S. Rijhwani, P. Riley, E. Salesky, F. Trabelsi, S. Winkler, B. Zhang, and M. Freitag. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284.
- [5] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat. MADLAD-400: A Multilingual and Document-Level Large Audited Dataset. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Datasets and Benchmarks Track.
- [6] M. Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, 2018, pages 186–191.
- [7] M. Popović. chrF: Character n-gram F-score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pages 392–395.
- [8] M. Popović. chrF++: Words Helping Character n-grams. In *Proceedings of the Second Conference on Machine Translation*, 2017, pages 612–618.
- [9] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of EMNLP 2020*, pages 2685–2702.
- [10] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, 2022.
- [11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL 2020*, pages 8440–8451.
- [12] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv:1907.05019, 2019.
- [13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.
- [14] P. Koehn. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, pages 388–395.
- [15] J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of LREC 2012*, pages 2214–2218.
- [16] Z. Wang, Z. C. Lipton, and Y. Tsvetkov. On Negative Interference in Multilingual Models: Findings and a Meta-Learning Treatment. In *Proceedings of EMNLP 2020*, pages 4438–4450.
- [17] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [18] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins. Hallucinations in Large Multilingual

- Translation Models. Transactions of the Association for Computational Linguistics, 11:1500–1517, 2023.
- [19] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. R. Mortensen, N. A. Smith, and Y. Tsvetkov. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In Proceedings of EMNLP 2023, pages 9904–9923.
- [20] A. Petrov, E. La Malfa, P. H. S. Torr, and A. Bibi. Language Model Tokenizers Introduce Unfairness Between Languages. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pages 36963–36990.

Appendix A. Fixed Run Configuration

Table A1: Configuration fixed prior to data collection. Any deviation encountered during execution will be reported in the final version rather than silently applied.

Parameter	Value
Model	facebook/nllb-200-distilled-600M
Precision	float16 on GPU; float32 on CPU
Decoding	Beam search, num_beams = 4, do_sample = False
Maximum input tokens	256
Maximum generated tokens	256
Batch size	16 (length-sorted, dynamic padding, original order restored)
Random seed	20260919
Evaluation split	devtest (final); dev (development and threshold setting)
Sentences per direction	300, identical FLORES identifiers across all directions
Primary metric	chrF++ (sacreBLEU CHRF, word_order = 2)
Secondary metrics	spBLEU (flores200 tokenizer); COMET wmt22-comet-da where supported
Bootstrap	1,000 resamples, 95% percentile interval
Annotation sample	25 flagged + 25 unflagged per direction, pooled and shuffled

Appendix B. Worked Example with Synthetic Data

EVERY NUMBER IN THIS APPENDIX IS SYNTHETIC

These values were invented to demonstrate that the analysis pipeline runs end to end and to show the shape of the completed paper. They are not measurements, were not produced by any model, and must not be cited, quoted, or carried into the final manuscript. Running `run_baseline.py` against FLORES-200 devtest overwrites every one of them. This appendix is to be deleted before submission.

The synthetic values below were chosen to be internally consistent and broadly plausible given the published reference points of Table 12. They deliberately illustrate the pre-registered hypotheses holding, which is precisely why they cannot substitute for data: a demonstration constructed to match a prediction confirms nothing about the world. Their function is narrower, namely to verify that the metric code, aggregation code, bootstrap, and figure code all execute and produce well-formed output before a GPU session is spent.

Table B1 (synthetic): Dataset and tokenization profile. Note the intended reading: Japanese has the fewest characters but more tokens than English, separating script density from tokenizer coverage.

Language	Script	Tier	Mean chars	Mean tokens	Inflation
English	Latin	High	148.2	31.2	1.00
French	Latin	High	168.5	36.8	1.18
Japanese	Japanese	High	71.4	41.5	1.33
Hindi	Devanagari	Medium	152.7	38.4	1.23
Bengali	Bengali	Medium	149.1	42.7	1.37
Swahili	Latin	Medium	157.3	45.1	1.45
Sinhala	Sinhala	Low	144.6	58.3	1.87
Yoruba	Latin	Low	181.9	61.9	1.98

Table B2 (synthetic): Main translation results. The into-English rows score above their out-of-English counterparts and the cross-pairs sit lowest, reproducing the pattern reported in [1].

Direction	Tier	chrF++ [95% CI]	spBLEU	COMET	s/sent
eng → fra	High	68.4 [66.9, 69.8]	44.2	0.861	0.21
eng → hin	Medium	52.1 [50.4, 53.7]	26.8	0.793	0.24
eng → ben	Medium	50.3 [48.6, 52.0]	24.1	0.780	0.25
eng → swh	Medium	51.7 [50.1, 53.4]	25.3	0.769	0.23
eng → sin	Low	44.8 [42.9, 46.6]	18.9	0.701	0.31
eng → yor	Low	38.2 [36.3, 40.1]	11.4	n/a	0.33
hin → eng	Medium	62.3 [60.8, 63.7]	35.7	0.842	0.22
swh → eng	Medium	59.8 [58.2, 61.3]	33.1	0.823	0.22
sin → eng	Low	54.1 [52.3, 55.8]	27.4	0.786	0.26
hin → swh	Cross	41.2 [39.4, 43.0]	15.2	0.690	0.27
swh → hin	Cross	39.6 [37.8, 41.4]	13.8	0.674	0.28

Table B3 (synthetic): Reliability flag rates, percentage of sentences. chrF++ falls by a factor of 1.8 from French to Yoruba (Table B2), while the off-target rate rises from 0.0 to 8.7 per cent and numeric corruption from 1.7 to 14.3 per cent — a proportionally far steeper gradient. This is the shape of result that would support H3.

Direction	Tier	Off-target	Src copy	Numeric	Repetition	Omis-susp.
eng → fra	High	0.0	0.0	1.7	0.3	2.3
eng → hin	Medium	0.3	0.3	3.3	0.7	5.0
eng → ben	Medium	0.7	0.3	4.0	1.0	5.7
eng → swh	Medium	1.0	1.3	4.7	1.3	6.7

Direction	Tier	Off-target	Src copy	Numeric	Repetition	Omis-susp.
eng → sin	Low	3.3	2.7	9.0	3.7	11.3
eng → yor	Low	8.7	7.3	14.3	6.0	17.0
hin → swh	Cross	6.0	3.0	10.7	4.3	13.7
swh → hin	Cross	5.3	2.7	11.3	4.7	14.0

Table B4 (synthetic): Fairness aggregates. System C raises the macro average by 0.5 and the worst-language score by 1.6. Under the rule fixed in Section 3.8, the macro difference would be reported as inconclusive and the worst-language difference examined against its own interval — an illustration of why the two columns are not interchangeable.

System	Worst chrF++	Worst language	Tier gap	SD	Macro chrF++
A — Baseline	38.2	Yoruba	26.9	9.4	51.1
B — English pivot	42.1	Swahili→Hindi	n/a	n/a	43.0
C — Constrained decoding	39.8	Yoruba	25.6	8.9	51.6

Table B5 (synthetic): Direct versus pivot. Quality rises and omission-suspect rate rises with it. This is the divergence H4 predicts, and it is invisible to any analysis reporting chrF++ alone — which is the methodological point of the protocol.

Pair	chrF++ direct	chrF++ pivot	Δ [95% CI]	Omis. direct	Omis. pivot
hin → swh	41.2	43.8	+2.6 [+0.9, +4.3]	13.7%	19.3%
swh → hin	39.6	42.1	+2.5 [+0.7, +4.2]	14.0%	20.7%

Table B6 (synthetic): Detector validation. Two detectors pass the pre-registered 0.80 precision threshold and two do not. This is the expected and desirable outcome: a validation step passing everything would indicate the threshold was set too low to discriminate.

Detector	Human label	Precision	Recall	n	Verdict under H6
Off-target	Wrong language	0.94	0.88	550	Passes — report as measured rate
Numeric error	Numeric error	0.87	0.91	550	Passes — report as measured rate
Repetition	Hallucination	0.61	0.43	550	Fails — screening only
Length anomaly	Omission	0.52	0.67	550	Fails — screening only

Reproducing this appendix: the repository accompanying this report includes `make_synthetic_demo.py`, which generates these tables from a synthetic translations file, exercising the same metric, aggregation, bootstrap, and plotting code paths used for real data. A change to the analysis code that breaks the pipeline will therefore surface here rather than mid-experiment.