

Sambalpuri Neural Translation: Data Refinement and Continued Fine-Tuning from V1 to V2

Rudra Narayan Meher

Undergraduate student, BS in Data Science and Applications
Indian Institute of Technology Madras (IIT Madras)
Research focus: Artificial Intelligence and Machine Learning

Abstract

This report studies continued fine-tuning for English-to-Sambalpuri neural machine translation using a mixed package of 6,400 records: 3,000 revised V1 examples, 2,800 new sentences, and 600 dictionary entries. Exact deduplication retains 6,335 records; source-grouped partitioning yields 5,994 training records and 331 fresh validation records, with 10 overlapping legacy records excluded. Starting from the saved V1 checkpoint, V2 is trained for five epochs on one NVIDIA A100-SXM4 GPU with 40 GB VRAM. Training completes in 1,037.43 seconds, and validation loss decreases from 1.159955 to 0.956640. On the author-corrected final benchmark of 149 unique reference pairs, BLEU increases from **6.78 to 33.54**, chrF++ from 39.69 to 58.91, and TER decreases from 78.15 to 49.67. A separate 331-record validation evaluation also favors V2, particularly on sentences. These findings support improvement over the project baseline within the supplied evaluations; they do not establish performance on an independently verified unseen test set.

Keywords: Sambalpuri; low-resource machine translation; corpus refinement; multilingual transfer; continued fine-tuning; SacreBLEU.

1. Introduction

The project targets translation from English into Sambalpuri, with target references written in Odia script. Its practical focus is a regional translation resource built through corpus collection, correction, and supervised adaptation. Multilingual transfer provides a relevant foundation [1], but the quality of aligned examples and the credibility of evaluation remain central to judging a small translation system.

The research question is whether a targeted update combining corrected legacy examples, new sentence coverage, and lexical entries improves the saved V1 system. The comparison is within the same project; it is not a leaderboard comparison against unrelated language pairs or test sets.

Contributions

(1) A documented 6,400-record update with auditable cleaning and split counts. **(2)** A five-epoch A100 training run with explicit optimization settings and compute measurements. **(3)** A corrected final benchmark, a separate validation comparison, and an evidence-based account of remaining lexical and evaluation limitations.

2. Corpus refinement and version history

The supplied V1 base manuscript and its revised version describe the earlier project. They are historical documentation of **V1**, not separate V2 experiments. The current V2 run continues training from a saved V1 checkpoint using the update package below. The earlier manuscripts report 15,000 parallel pairs; those historical pairs were not supplied for a record-level audit, and that number is not added to 6,400 as a verified unique corpus total. [S1-S4]

2.1. What the 6,400-record update contains

Component	Raw rows	Share	Role
Revised V1 examples	3,000	46.875%	Existing examples with issue or grammar fixes
New sentence pairs	2,800	43.750%	Additional sentence-level coverage
Dictionary / word entries	600	9.375%	Lexical supervision
Total update package	6,400	100%	3,400 newly added records; 3,000 revised records

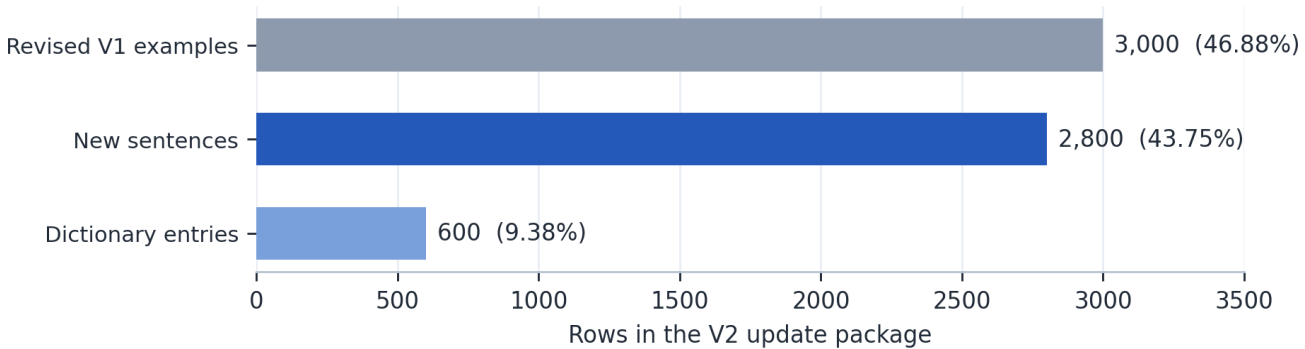


Figure 1. Composition of the V2 update package. The package contains 6,400 rows, but only 3,400 are newly added sentences or dictionary entries; the remaining 3,000 revise legacy material.

2.2. Cleaning and leakage controls in the notebook

The notebook strips blank fields and normalizes source and target text using Unicode NFC, collapsed whitespace, and case folding. It removes exact duplicates using normalized source, normalized target, and record type, while preserving different target translations. It flags **277 English sources with multiple target translations** for review; this flag can represent valid variation or an alignment issue and is not treated as 277 confirmed errors. [S3, S5]

Origin	Raw	Unique	Train	Validation	Excluded
Revised V1	3,000	2,993	2,983	0	10
New sentences	2,800	2,742	2,462	280	0
Dictionary entries	600	600	549	51	0
Total	6,400	6,335	5,994	331	10

Table 1. Reconciled V2 record counts. Unique new-sentence counts are derived from the saved origin totals. The 65 removed duplicates equal 1.02% of the raw package. The 10 excluded legacy records share normalized English sources with validation records.

GroupShuffleSplit uses a 10% group holdout and seed 42 on fresh records. English sources stay together, and legacy records overlapping those validation sources are excluded from training. The executed assertion confirms no normalized-source overlap between the current training and validation frames. It does not independently audit the full historical V1 corpus.

3. Model and experimental methodology

3.1. Continued fine-tuning and language routing

V2 loads the saved V1 model through Hugging Face AutoModelForSeq2SeqLM and continues supervised training. The project identifies this as an NLLB-200-based translation setup [1, 6]; the executed checkpoint reports `model_type = m2m_100`. The original upstream model identifier, immutable revision, and parameter count are not available in the supplied evidence, so no exact model-size claim is made.

The tokenizer uses `eng_Latn` for English and `ory_Orya` as the target language identifier, with target BOS token ID `256136`. Thus, Sambalpuri adaptation is learned from project examples while generation is routed through the Odia language token. This setup does not introduce a dedicated Sambalpuri language token. NLLB-200 is a model-family name, not evidence of a 200-million-parameter checkpoint. [S3]

3.2. Recorded training configuration

Setting	Executed configuration
Initialization and objective	Saved V1 checkpoint; supervised seq2seq token prediction
Hardware / precision	1 x NVIDIA A100-SXM4-40GB; BF16 enabled; FP16 disabled
Training / validation records	5,994 / 331
Epochs / optimizer updates	5 completed / 1,875
Batch size / accumulation	4 examples per device x 4 accumulation steps; effective batch 16
Optimizer / initial learning rate	AdamW (adamw_torch) / 0.00001
Schedule / warmup	Linear decay / 94 optimizer updates (approximately 5%)
Weight decay / gradient clipping	0.01 / maximum gradient norm 1.0
Maximum source / target length	128 / 128 tokenizer tokens
Seeds / checkpoint selection	42 / lowest validation loss; evaluated and saved each epoch
Early stopping / checkpointing	Patience 2; all five epochs completed; gradient checkpointing off
Logged software	Transformers 5.20.2; SacreBLEU 2.6.0; Python 3.13.16

3.3. Evaluation protocol

Both checkpoints are decoded with batch size 8, beam size 4, a 128-token source limit, and up to 128 newly generated tokens. The final CSV benchmark reports corpus BLEU [2, 3], chrF++ with word_order=2 [4], TER [5], exact match after edge-whitespace stripping, inference time, and examples per second. Timing begins after model loading and includes tokenization, generation, and decoding, with CUDA synchronization around the measured interval. [S3]

The separate 331-record validation analysis defines exact match using NFC, whitespace normalization, and case folding. Its exact-match percentages therefore have a different normalization rule from the final CSV benchmark. SacreBLEU signatures are not printed in the supplied run, so the report records the package version and metric calls rather than inventing a full signature.

4. Training behavior and compute profile

The V2 run completes all five scheduled epochs and selects **checkpoint-1875** by minimum validation loss. Validation loss decreases at every recorded epoch, from 1.159955 to 0.956640, a **17.53% relative reduction**. The logged training value rises slightly in the final epoch, so its trajectory is not described as monotonic. [S3-S7]

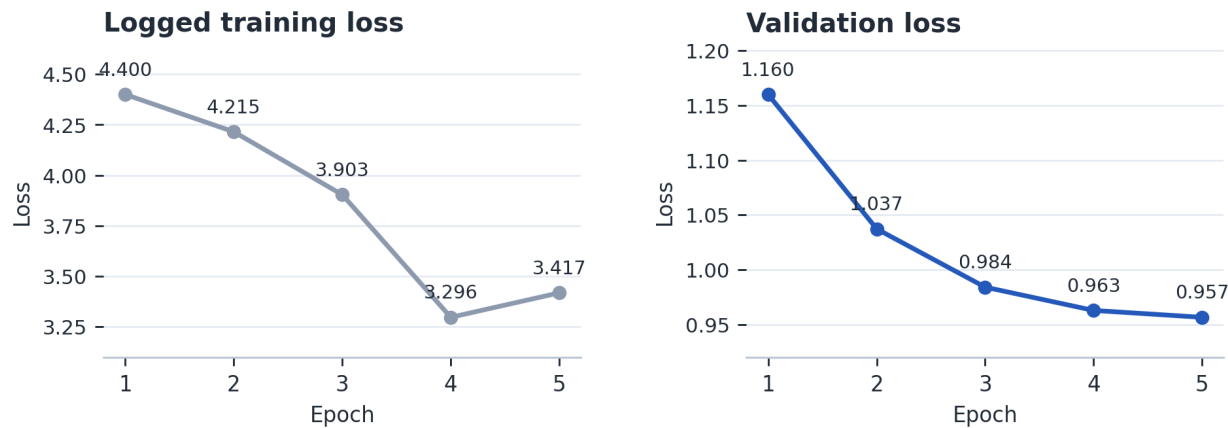


Figure 2. V2 loss values from the notebook Trainer table, corroborated by the supplied training screenshot. Separate y-axes are used. Training values are the Trainer-displayed logging values at epoch boundaries, not independently recomputed full-epoch means.

Epoch	Logged training loss	Validation loss
1	4.399788	1.159955
2	4.215046	1.037327
3	3.903251	0.984181
4	3.296158	0.962923
5	3.417338	0.956640

Compute measurement	Recorded value
Training wall time	1,037.43 seconds (17.29 minutes)
Optimizer update throughput	1.807 steps/s
Training example throughput	28.889 examples/s
Run-averaged training loss	4.143593
Best validation loss	0.956640
Peak allocated GPU memory	12.856 GiB, as logged by torch.cuda.max_memory_allocated
GPU capacity	40 GiB physical VRAM (42,949,672,960 bytes in metadata)

4.1. Interpretation

The results show successful optimization of the continued-training objective with practical single-GPU runtime. The displayed training loss and validation loss use different collection conditions and should not be read as a calibrated generalization-gap estimate. Checkpoint selection uses the same 331 examples later scored as validation; those scores measure development performance. They are not a substitute for a locked test set.

5. Final author-corrected benchmark

The final benchmark uses **149 unique English-reference pairs**. The author identifies the corrected figure, `sambalpuri_benchmark_corrected.png`, as the final result and explicitly corrects V1 BLEU to **6.78**. This report follows that correction consistently. All other final values below agree with the supplied figure and the saved benchmark display. [S3, S8, S9]

Metric	V1	V2	Change (V2 - V1)
BLEU (higher is better)	6.78	33.54	+26.76 points
chrF++ (higher is better)	39.69	58.91	+19.22 points
TER (lower is better)	78.15	49.67	-28.48 points
Exact match	5.37%	17.45%	+12.08 percentage points
Inference time	7.13 s	5.53 s	-1.60 s
Throughput (examples/s)	20.89	26.96	+6.07 examples/s

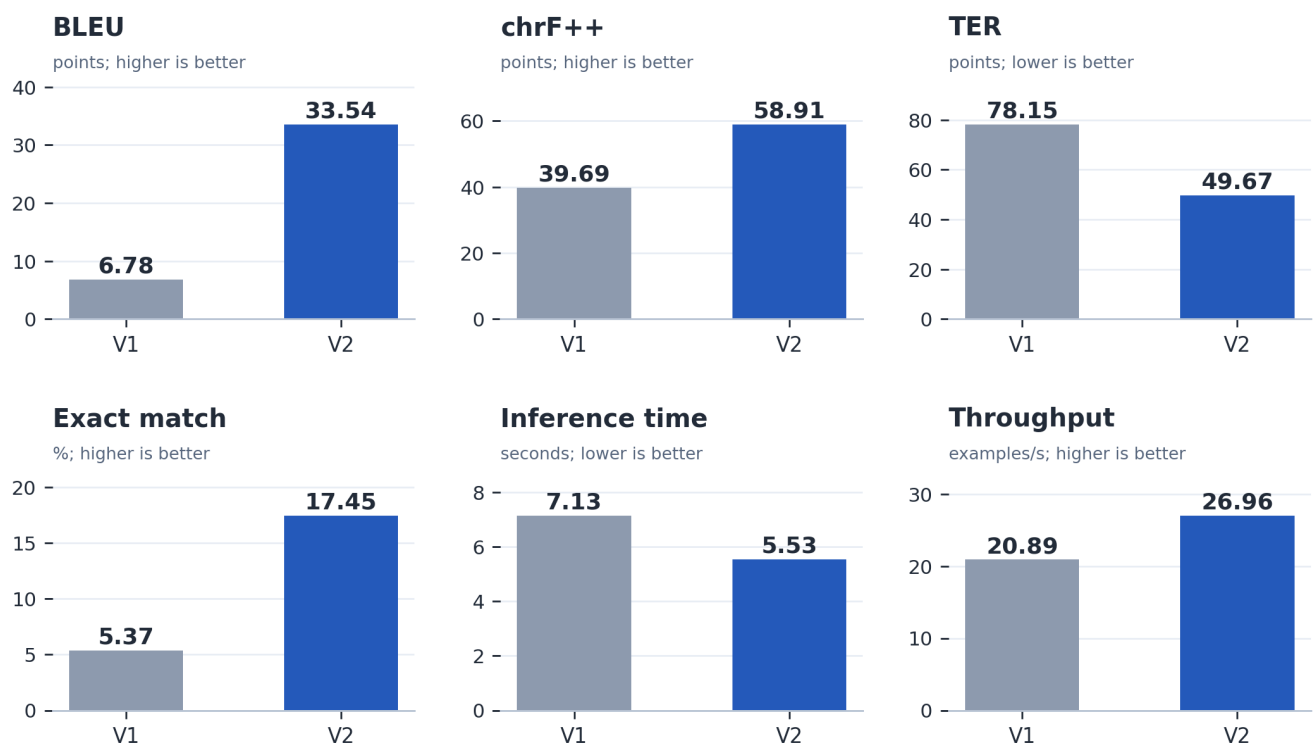


Figure 3. Final corrected V1-to-V2 comparison. V1 is gray; V2 is blue. Each panel has its own unit and scale. Exact match corresponds to 8/149 pairs for V1 and 26/149 for V2, inferred from the reported percentages.

5.1. What the improvement establishes

V2 improves across all six reported measures on this benchmark. Lower TER indicates fewer reference edits under the metric, while higher chrF++ indicates closer character and word n-gram agreement. BLEU 33.54 is an overlap score, not 33.54% translation accuracy. These values support improvement on the measured sample, without establishing a universal quality rating or superiority over unrelated published systems. [2-5]

Correction and timing note. V1 BLEU is an author-supplied correction that supersedes an earlier notebook display. It has not been recomputed here from a complete prediction export. The timing comparison is one recorded run per checkpoint, excludes model loading, and has no repeat-run variance estimate. The final set has not been independently certified as unseen to both V1 and V2.

6. Separate validation evaluation: 331 records

The training notebook also compares V1 and V2 on the **331 fresh validation records**: 280 sentences and 51 dictionary entries. These are the same records used for validation-loss checkpoint selection. They are a different population from the final 149-pair CSV benchmark and must retain their own scores and denominator. [S3-S5]

Subset / model	n	BLEU	chrF++	Normalized exact match
All / V1	331	7.27	35.58	2.11% (7/331)
All / V2	331	22.18	49.93	7.25% (24/331)
Sentences / V1	280	7.41	36.10	0.00% (0/280)
Sentences / V2	280	22.36	50.69	4.29% (12/280)
Words / V1	51	2.88*	19.17	13.73% (7/51)
Words / V2	51	0.00*	20.98	23.53% (12/51)

Table 2. Validation results from the machine-readable W&B summary. *Corpus BLEU on a word-only subset can become zero because standard higher-order word n-grams are unavailable. This zero should not be interpreted as zero lexical accuracy; chrF++ and exact match provide more suitable lexical diagnostics.

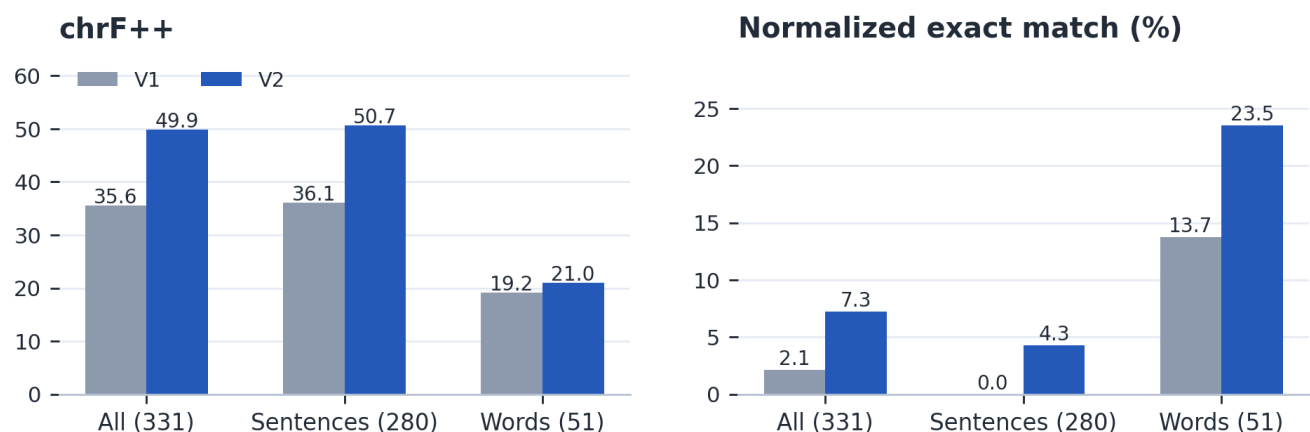


Figure 4. Validation subgroup comparison. V2 shows a larger character-overlap improvement on sentences (+14.58 chrF++) than on words (+1.82 chrF++). Dictionary translation remains a weaker part of the system.

6.1. Exact-match transitions

V1 status	V2 not exact	V2 exact	Total
Not exact	306	18 newly exact	324
Exact	1 regression	6 retained	7
Total	307	24	331

The 18 corrected examples and one regression produce a net gain of 17 exact matches. Twelve corrections occur among sentences; six corrections and the one regression occur among word entries. This shows that aggregate improvement coexists with individual failures. These are normalized string-match transitions, not counts of human-verified semantic corrections.

6.2. Why the two benchmark scores differ

The final V2 BLEU of 33.54 and validation BLEU of 22.18 are not conflicting measurements: they come from different example sets. Sample composition, length, reference wording, and potential training overlap can change scores. The supplied evidence does not isolate which factor explains their difference, so the report does not attribute the gap to a specific cause.

7. Benchmark audit and qualitative findings

7.1. Actual size and source diversity

The attached CSV has **150 rows and two columns**, English and Reference. It contains source-reference pairs, not saved model predictions, despite the filename containing `benchmark_500_predictions`. Applying the notebook's stripping and exact pair-deduplication rule removes one repeated pair, giving **149 usable pairs and 142 distinct English sources**. The repeated source-reference pair is "Arrange the bed." [S8]

Audit field	Observed result
Raw / deduplicated reference pairs	150 / 149
Distinct English sources / target references	142 / 149
Missing source or reference fields	0
Mean / median English source length	4.54 / 5 whitespace tokens
Source length range / single-token sources	1-11 tokens / 19 pairs

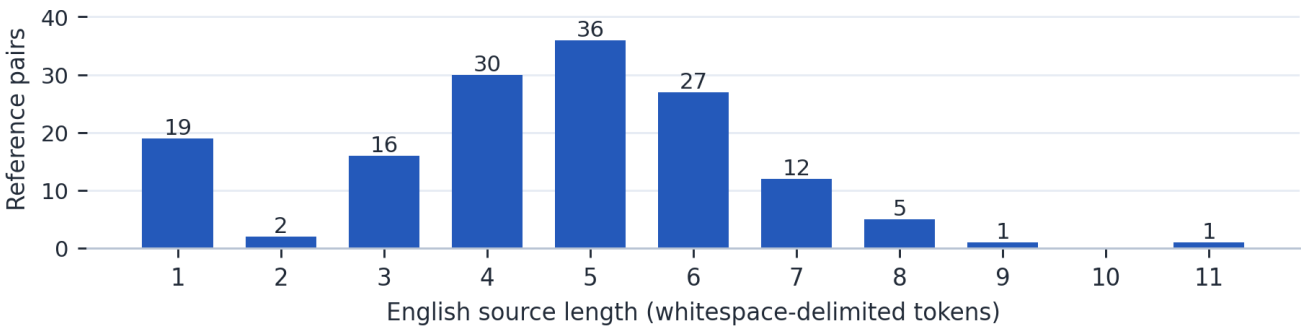


Figure 5. English source-length distribution across the 149 retained reference pairs, computed directly from the attached CSV. Short prompts dominate; these results do not establish paragraph-level or long-context translation performance.

Some English sources have more than one reference row, so the set has 149 distinct pairs rather than 149 independent English prompts. A future uncertainty analysis should resample at the source-group level. The supplied training CSV and full historical V1 corpus are absent, preventing a complete train-test overlap audit for this final set.

7.2. Examples visible in the saved notebook output

English input	Observed finding	Implication
Papaya	The V2 output names onion rather than the reference fruit.	A concrete lexical substitution remains.
Your (Informal)	V1 retains the English annotation; V2 matches the reference without it.	Less annotation leakage in this displayed example.
Mita's phone is charging.	The displayed V2 output exactly matches its reference; V1 does not.	A successful local example, not a success-rate estimate.

Table 3. Selected observations from the final comparison display. Target strings are described in English. These examples are illustrative and were not drawn as a blinded random human-evaluation sample. [S3]

Spelling, spacing, and alternative valid wording can reduce overlap even when a translation is understandable. Conversely, overlap can remain high despite a critical meaning error. A bilingual adequacy and fluency study is needed to distinguish these cases systematically.

8. Discussion, limitations, and next experiments

8.1. Supported interpretation

The strongest result is a consistent V1-to-V2 improvement across the final author-corrected benchmark and the separate development evaluation. The work demonstrates an implemented adaptation workflow: assemble and revise data, audit duplicates and source overlap, continue training, select a checkpoint by validation loss, and compare decoded outputs. The supplied evidence supports the utility of the combined update; it does not identify the independent contribution of corrections, new sentences, or dictionary entries.

8.2. Limits that affect scientific interpretation

Evaluation independence. The 331-record set is used in checkpoint selection, and complete historical V1 exposure is unknown. The final 149-pair set has no independently verified overlap audit. Neither set can be presented as a certified unseen test set on current evidence.

Reproducibility and uncertainty. Full final prediction exports, immutable checkpoint hashes, the upstream model revision, complete training manifests, and printed metric signatures are absent from this bundle. The final V1 BLEU follows the author's explicit correction. No repeated seeds, paired confidence intervals, or repeated timing trials are reported.

Coverage and causality. The final examples are short, and lexical errors remain. No ablation separates revised legacy examples from new sentence or word supervision. No comparison against an unadapted upstream checkpoint, another multilingual baseline, or a blinded bilingual evaluation is available.

Historical and release scope. Earlier corpus totals and loss values are retained as historical context, not merged into V2 results. Pair-level permissions and privacy review for historical web or conversation data are not established by the training metrics and should accompany any dataset release.

8.3. Prioritized experimental follow-up (not yet executed)

Experiment	What it would resolve
Lock a source-grouped test set	Deduplicate against all V1/V2 material and freeze it before further tuning.
Ablate the update components	Compare revised-only, new-sentence-only, and combined training under fixed settings.
Evaluate lexical and longer inputs	Separate word adequacy, sentence translation, and longer-passage behavior.
Repeat and publish the comparison	Export full predictions and metric signatures; repeat seeds and timed inference; obtain blinded bilingual ratings.

9. Conclusion

Continued fine-tuning on the documented update produces a stronger Sambalpurī translation checkpoint within the supplied comparisons. The final corrected benchmark records BLEU **6.78 to 33.54**, chrF++ **39.69 to 58.91**, and TER **78.15 to 49.67**. The next scientific milestone is to establish the same gain on a locked, independently audited test set with complete prediction and checkpoint provenance.

Author contributions and acknowledgment

Rudra Narayan Meher: model development, training, evaluation, analysis, and research reporting.

Acknowledgment: Shivangi Dwivedi contributed to dataset collection only.

Appendix A. Historical V1 documentation

Both supplied V1 PDFs concern the earlier study. The base manuscript is dated 10 December 2025; the revised manuscript explicitly revisits that original report. Their descriptions and loss scales are retained here to preserve project history. They do not describe the A100 V2 run, and their loss values are not directly comparable with V2 losses because the exact earlier checkpoint and loss-reduction details are unspecified. [S1, S2]

A.1. Corpus composition reported for the earlier study

Reported source	Pairs	Share
Regional social media captions	5,000	33.3%
Manually curated translations	5,000	33.3%
Internet comments / forums	3,000	20.0%
Anonymized conversations	2,000	13.3%
Total reported in V1 manuscripts	15,000	100%

Table A1. Historical source counts reproduced from the V1 manuscripts. The historical pair-level corpus is not included in this evidence bundle. These counts do not establish the net unique training exposure of the saved V1 checkpoint or the cumulative V2 corpus size.

A.2. Historical training trace

Recorded epoch	Training loss	Validation loss
1	0.0590	0.0498
2	0.0440	0.0408
3	0.0346	0.0375
4	0.0215	0.0355
5	0.0212	0.0345
6	0.0165	0.0341

The old manuscripts describe a Tesla T4 with 16 GB VRAM. Some prose says seven epochs, but only six numeric epochs are provided; the table above therefore contains six records. Historical training loss declines by 72.03% and validation loss by 31.53% across those six rows. These reductions concern the historical objective and do not establish a translation-quality score.

A.3. How to read the versions

Evidence stage	Use in this report
Original V1 manuscript	Historical corpus description, T4 context, and six-row loss series
Revised V1 manuscript	Clarified interpretation of the same earlier experiment
V2 notebook and A100 run	Current configuration, data cleaning, five-epoch training, and 331-record validation
Final corrected comparison figure	Authoritative reported 149-pair final benchmark, including V1 BLEU 6.78

No historical qualitative accuracy percentage or claims of state-of-the-art performance are carried into the updated results. The current measured contribution is the within-project checkpoint comparison documented in Sections 4-7.

References and evidence provenance

[1] NLLB Team et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672. <https://arxiv.org/abs/2207.04672>

[2] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. ACL, pp. 311-318. <https://aclanthology.org/P02-1040/>

[3] Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. WMT, pp. 186-191. <https://aclanthology.org/W18-6319/>

[4] Popovic, M. (2017). chrF++: words helping character n-grams. WMT, pp. 612-618. <https://aclanthology.org/W17-4770/>

[5] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. AMTA, pp. 223-231. <https://aclanthology.org/2006.amta-papers.25/>

[6] Hugging Face. NLLB model documentation. Accessed 24 October 2026. https://huggingface.co/docs/transformers/en/model_doc/nllb

Appendix B. Source map

ID	Supplied evidence	Claims supported
S1	smb1_v1_base (1).pdf	Original V1 history, reported corpus and loss trace
S2	smb1_v1_revise(1).pdf	Revised interpretation of the same V1 study
S3	Copy_of_sambalpuriv2.ipynb	Executed code, counts, settings, outputs and final chart cell
S4	output.log	Training summary and 331-record evaluation scalars
S5	wandb-summary.json	Data counts, validation metrics, runtime and peak allocation
S6	wandb-metadata.json	A100-SXM4-40GB hardware and Python version
S7	image(20261024-134837).png	Five-epoch training and validation loss display
S8	150_branhmark - benchmark_500_predictions(1).csv	150 reference pairs; 149 after exact pair deduplication
S9	image(20261024-134831).png	Final corrected graph; also reproduced by notebook cell 31
S10	Screenshot at 7.20.10 PM	Visual W&B training, learning-rate and gradient trends

Source precedence. The author's correction and final graph govern the final V1 BLEU value. Executed notebook outputs and machine-readable scalars govern V2 training and validation. The two V1 manuscripts are historical evidence only. Metric deltas and the final CSV audit are calculated in preparing this report; model training and generation were not rerun.

The W&B screenshot shows an initial loss decline and a warmup followed by learning-rate decay. The summary JSON contains terminal scalars and gradient histograms, not a complete numeric step history. The report therefore replots the five recorded epoch values without reconstructing unsupported step-level curves.

Attached reference CSV SHA-256:
34018ded6ee4e31c3c3e7c2bade30794
51e6d9edba56e24eae69b4877e9d3385

Prepared from the supplied evidence as of 24 October 2026.