

SCE-KIIT: KokLLaMA: Cross-Task Adaptation for Low-Resource English–Kokborok Translation

Agniva Maiti[†] Aarsh Muley[†] Sovan Kumar Sahoo[†]

[†]School of Computer Science and Engineering
Kalinga Institute of Industrial Technology (KIIT) Deemed to be University
Bhubaneswar-751024, Odisha, India

2205964@kiit.ac.in 22051304@kiit.ac.in sovan.sahoofcs@kiit.ac.in

Abstract

Kokborok is a critically under-resourced Tibeto-Burman language spoken by over one million people primarily in Tripura, India, and is absent from modern massively multilingual machine translation (MT) systems. We describe the SCE-KIIT submission to the WMT 2026 Shared Task on Low-Resource Indic Language Translation (Category 2: English $\leftrightarrow$ Kokborok). We ask whether a large language model can be adapted to translate a language using *no parallel supervision at all*. Our system, KokLLaMA-3.2-3B-Instruct, fine-tunes Llama-3.2-3B-Instruct via QLoRA purely on Kokborok conversational instruction data, and recovers translation behaviour at inference time through structured prompting and a rule-based post-processing pipeline. On the official WMT 2026 test set it obtains 5.11 BLEU / 27.09 chrF++ (EN $\rightarrow$ TRP) and 3.58 BLEU / 26.14 chrF++ (TRP $\rightarrow$ EN), placing third of four and third of five submitted primary systems, and coming within 1.9 BLEU of the best WMT 2025 system for this pair despite using no parallel data. The approach is substantially more effective at generating Kokborok than at generating English, and we analyse why. We further show that our in-domain validation split under-estimated official performance by more than an order of magnitude, a cautionary result for participants who tune on held-out slices of provided training corpora, and we characterise a pathological Translation Edit Rate (TER) failure mode in which an untrained baseline scores well by silently under-generating.

1 Introduction

Recent years have seen rapid advancements in neural machine translation (NMT), largely driven by the scaling of sequence-to-sequence models and Large Language Models (LLMs) trained on massive web-scale corpora. However, the data-hungry nature of these models creates a digital divide for low-resource languages that lack substantial web

representation. Minor, regional, and indigenous languages spoken by smaller communities are frequently excluded from modern multilingual MT frameworks, leaving them digitally marginalized (Ranathunga et al., 2023; Ebrahimi et al., 2022).

Kokborok (also known as Tripuri) is one such language. It is a Tibeto-Burman language native to the state of Tripura in northeastern India and neighboring regions of Bangladesh, spoken as a mother tongue by over one million people. While historically written in various scripts, written Kokborok has increasingly adopted the Latin (Roman) script in recent decades. Despite its vital sociolinguistic role and official state recognition, digital resources for Kokborok are sparse. It is entirely absent from major multilingual models like NLLB-200 (Costa-jussà et al., 2022) and has no dedicated translation system, which poses a severe challenge for standard MT transfer methods.

To address this gap, we participate in the WMT 2026 Low-Resource Indic Language Translation Shared Task (Category 2: English $\leftrightarrow$ Kokborok). In this paper, we describe KokLLaMA-3.2-3B-Instruct, our submission built upon the Llama-3.2-3B-Instruct architecture (Dubey et al., 2024). Our approach is centered on instruction tuning: we leverage a curated conversational instruction corpus of 11,428 QA pairs in Kokborok to teach the model lexical and grammatical patterns in the target language. At inference time, we translate using a highly structured prompt designed to force translation behavior, combined with a rule-based post-processing cleanup script.

This paper makes four contributions. First, we show that cross-task adaptation alone is competitive in this setting: on the official WMT 2026 test set the system reaches 5.11 BLEU (EN $\rightarrow$ TRP) and 3.58 BLEU (TRP $\rightarrow$ EN) without observing a single parallel sentence pair during training, placing third of four and third of five submitted primary systems and coming within 1.9 BLEU of

the best WMT 2025 system for this pair. Second, we identify a pronounced directional asymmetry – the approach transfers far better when generating Kokborok than when generating English – and trace it to the generative direction of the instruction corpus. Third, we show that our internal validation split, drawn from the biblical-domain parallel corpus, *under*-estimated official test performance by more than an order of magnitude (0.37 vs. 5.11 BLEU for EN→TRP), a cautionary result for participants who tune on held-out slices of provided training data. Fourth, we analyse a pathological failure mode in the Translation Edit Rate (TER) metric (Snover et al., 2006): the unmodified base model achieves a misleadingly “good” TER (≈ 99.45) by failing to translate at all and emitting empty or very short strings (the 100-TER cap), whereas the fine-tuned KokLLaMA model attempts translation but suffers EOS-token generation failures that inflate TER beyond 200. Our post-processing pipeline reduces TER substantially in-domain (from 203.55 to 108.06 for EN→TRP), though the official TRP→EN score of 200.86 shows that the problem persists out-of-domain.

2 Related Work

2.1 Low-Resource Machine Translation

Translation in data-scarce environments has evolved from phrase-based statistical systems (SMT) (Koehn et al., 2007) to deep transformer-based architectures (Vaswani et al., 2017). Common strategies to mitigate data scarcity include subword vocabulary engineering via Byte-Pair Encoding (BPE) (Sennrich et al., 2016), cross-lingual transfer from high-resource typological relatives (Zoph et al., 2016; Neubig and Hu, 2018), and back-translation data augmentation (Edunov et al., 2018). Combining massively multilingual pre-trained models with target-language instruction fine-tuning is the current state-of-the-art low-resource paradigm (Ebrahimi et al., 2022; Ranathunga et al., 2023).

2.2 Instruction Tuning of LLMs

Large Language Models (LLMs) have shown strong translation capabilities when prompted appropriately, often matching or exceeding supervised NMT models in high-resource scenarios (Brown et al., 2020). For low-resource languages, however, vanilla prompting fails due to the model’s

limited pre-training exposure. Instruction fine-tuning via supervised instruction adaptation (SFT) has emerged as a key method to adapt LLMs to unrepresented target languages, enabling zero-shot and few-shot cross-task transfer from conversational QA to structured MT tasks.

2.3 Indic Machine Translation

Research in Indic MT has historically focused on the 22 constitutionally scheduled languages, yielding massive parallel databases like Samanantar (Ramesh et al., 2022) and IndicNLP Suite benchmarks (Kakwani et al., 2020). The WMT Low-Resource Indic Language Translation Shared Tasks have expanded this coverage to unrepresented regional languages, with prior editions addressing English–Nyishi (Kakum et al., 2023) and other minority pairs (Pal et al., 2023; Pakray et al., 2024, 2025). English–Kokborok is *not* newly introduced at WMT 2026: the pair was already evaluated under Category 2 at WMT 2025 (Pakray et al., 2025), where the best submitted system reached 6.99 BLEU (EN→TRP) and 2.99 BLEU (TRP→EN). We therefore position this work as a contribution to an existing benchmark rather than as its first evaluation, and compare our official scores against those figures in Section 6.1.

3 Dataset and Preprocessing

3.1 Instruction Tuning Data

For SFT fine-tuning we use a Kokborok instruction-following corpus of 11,428 conversational question–answer pairs, written in the Roman script and covering everyday conversation, elementary STEM material, and cultural knowledge specific to Tripura. Under the 95/5 split reported in Table 1 this yields 10,856 training pairs and 572 held-out pairs. The corpus allows the model to acquire the orthography, syntax, and morphology of Kokborok even though the training objective is conversational rather than translational.

Released version. For reproducibility we state the relationship between this corpus and its public release precisely. The released dataset `agnivamaiti/kokborok-qa` is an earlier and smaller snapshot: it contains 4,943 rows, of which 2,514 are distinct prompt–completion pairs, and 81.7% of those distinct pairs also occur in the 11,428-pair corpus used for training. It is therefore neither a deduplicated copy nor a strict subset of the corpus we train on. All experiments reported

here use the 11,428-pair version, and a re-run using only the released snapshot will not reproduce our numbers.

3.2 Parallel Evaluation Data

For evaluation, we utilize the official WMT 2026 English–Kokborok training parallel corpus (English–Kokborok Training Data 2026.xlsx), which contains 2,266 high-quality aligned pairs. To ensure full experimental reproducibility, we split the data deterministically using a fixed random seed (`random_state=42`):

- **Train Set:** 1,812 sentence pairs (80%) reserved for few-shot prompt construction.
- **Validation Set:** 227 sentence pairs (10%) used for prompt validation.
- **Held-out Evaluation Set:** 227 sentence pairs (10%) reserved for internal model evaluation.

Domain mismatch. We stress that this held-out set is *not* the official test set. It is a 10% slice of the WMT 2026 *training* corpus, whose content is drawn predominantly from biblical text. The official WMT 2026 test set, by contrast, consists of 1,000 news-domain sentences per direction and is distributed without references, so participants cannot score it locally. Every number we report on $n = 227$ is therefore an in-domain validation figure, and Section 6.1 shows that these figures diverge sharply from the official results.

We also apply a basic cleaning pipeline. Sentences are converted to Unicode NFC form. Quotes and apostrophes are replaced with ASCII equivalents to reduce tokenizer fragmentation. Length outliers (fewer than 3 or more than 200 words) and asymmetric lengths (ratio exceeding 5.0) are removed to eliminate noisy alignments.

4 Methodology

4.1 Model Fine-Tuning

We build our system on Llama-3.2-3B-Instruct. The model is adapted with QLoRA (Hu et al., 2021; Dettmers et al., 2023) under NF4 4-bit quantization with double quantization and bfloat16 compute dtype via `bitsandbytes`, using the `SFTTrainer` supervised fine-tuning loop. The full configuration is given in Table 1. The resulting adapter is released as `agnivamaiti/KokLLaMA-3.2-3B-Instruct` and the pipeline is illustrated in Figure 1(a).

Hyperparameter	Value
LoRA rank r	64
LoRA α	128
LoRA dropout	0.05
Target modules	<code>q, k, v, o, gate, up, down-proj</code>
Quantization	NF4, double quant, bf16 compute
Epochs	3
Per-device batch size	8
Gradient accumulation	2 (effective batch 16)
Learning rate	2×10^{-4}
Optimizer	<code>paged_adamw_32bit</code>
Max sequence length	1,024
Train / eval split	95 / 5

Table 1: QLoRA fine-tuning hyperparameters for KokLLaMA-3.2-3B-Instruct.

Unrecorded settings. The warmup schedule, hardware configuration, and wall-clock training time were not logged during the original fine-tuning run. We report this omission rather than reconstruct the values post hoc.

4.2 Translation Prompting

To align the conversational model for machine translation we construct structured prompts. For English→Kokborok the prompt is, verbatim:

```
User: Translate this English
sentence into Kokborok (Roman
script).
Rules: Output ONLY the Kokborok
translation. No English words.
No explanations. No notes. No
parentheses. No quotes.
English: {sentence}
Assistant: Kokborok:
```

and for Kokborok→English:

```
User: The sentence below
is written in Kokborok, a
Tibeto-Burman language spoken
in Tripura, India. What does it
mean in English?
Rules: Write only a fluent
English sentence. Do not copy
Kokborok words. No explanations.
No notes.
Kokborok: {sentence}
Assistant: In English, this
means:
```

Both prompts terminate in a completion primer (Kokborok: and In English, this means:) rather than in an instruction. We found this to be the strongest available mechanism for suppressing source copying without retraining, since it forces the model to continue in the target language rather than echo the input. The zero-shot baseline of Section 5.2 uses the same structure, expanded only to gloss Kokborok as “a

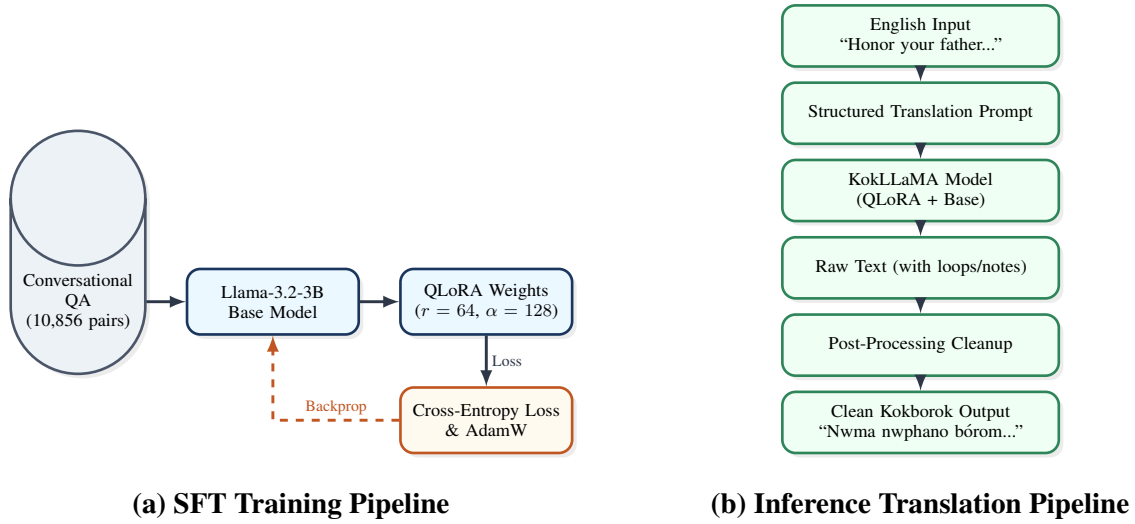


Figure 1: System architecture flowchart. (a) SFT training pipeline; (b) Inference translation pipeline using structured prompts and rule-based post-processing.

Tibeto-Burman language of Tripura, India, written in Roman script”, as the base model has no reliable association for the language name alone.

4.3 Rule-Based Post-Processing

Because KokLLaMA is fine-tuned on a conversational QA dataset, it frequently displays conversational behavior during inference. Common issues include:

- Emitting conversational preambles (e.g., “Sure, here is the translation:”) or explanatory notes in parentheses (e.g., “(Note: Kaitor means Lord in Kokborok)”).
- Failure to output the end-of-sequence (EOS) token, leading to infinite repetition loops of target words or prompt sections.

To mitigate this we apply a deterministic rule-based function to every generated string. Generation is first truncated at the first newline; the function then performs six steps in order:

1. Strip the completion primer if the model echoes it back (Kokborok: or In English, this means:).
2. Strip a fixed set of explanatory prefixes observed in the SFT outputs (Answer:, Translation:, English meaning: and their Kokborok equivalents).
3. If the output is wrapped in quotation marks, extract the first quoted span.
4. Delete trailing parenthetical notes matching (Note: ...) and abbreviation expansions of the form (ABC = ...).

5. Strip residual leading and trailing quotation characters.
6. Split on sentence boundaries and retain only the first sentence.

We emphasise that the pipeline contains *no* explicit repetition or duplicate-phrase detector. Loop truncation is an emergent consequence of step 6 and of the first-newline cut: because degenerate repetitions almost always begin after the first complete sentence, retaining only that sentence removes them. This also explains the pipeline’s two known failure modes. It cannot remove *inline* parenthetical glosses such as (face) or (dust), which occur mid-sentence rather than as trailing note blocks, and it silently discards legitimate multi-sentence translations. The full inference pipeline is diagrammed in Figure 1(b).

5 Experimental Setup

5.1 Evaluation Metrics

We evaluate translation performance using three metrics:

1. **SacreBLEU** (Post, 2018): Word-level BLEU score with standard signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp`.
2. **chrF++** (Popović, 2015, 2017): Character n -gram F-score, which is robust to orthographic variations.
3. **TER** (Snover et al., 2006): Translation Edit

Dir	Run	BLEU	MET	chrF++	TER	COMET
EN→TRP	Primary (beam 4)	5.11	21.04	27.09	98.67	54.35
EN→TRP	Contrastive (greedy)	1.98	11.94	20.32	108.65	49.89
TRP→EN	Primary (beam 4)	3.58	23.59	26.14	200.86	53.49
TRP→EN	Contrastive (greedy)	1.45	16.72	22.13	137.37	53.31

Table 2: Official WMT 2026 results for SCE-KIIT on the blind test set ($n = 1,000$ per direction), as released by the organisers. BERTScore is omitted for space (71.17 / 68.62 for EN→TRP primary/contrastive; 86.99 / 87.00 for TRP→EN). Full figures appear in the findings paper (Pakray et al., 2026).

Rate (lower is better), measuring the edit actions required to match references.

5.2 Baseline Systems

We compare KokLLaMA against the unmodified zero-shot baseline of the base Llama-3.2-3B-Instruct model. The baseline is prompted identically, ensuring a clean comparison of fine-tuning gains.

5.3 Decoding Configuration

All systems are decoded with identical settings. Our primary run uses beam search with `num_beams=4` and `early_stopping=True`; the contrastive run is identical except for greedy decoding (`num_beams=1`). Sampling is disabled throughout (`do_sample=False`), so temperature and top- p are inert. We generate at most `max_new_tokens=128` with `repetition_penalty=1.3`, batch size 16 (32 for the contrastive run), left padding, and inputs truncated to 512 tokens. Inference is run on the 4-bit NF4 quantized model.

Notably, we do *not* set `no_repeat_ngram_size`. In hindsight this is the most likely proximate cause of the repetition loops analysed in Section 6.4: `repetition_penalty` rescales logits but does not hard-block repeated n -grams, so our post-processing pipeline was left to absorb failures that a decoding constraint could have prevented outright.

6 Results and Analysis

6.1 Official WMT 2026 Results

Table 2 reports the official scores released by the shared task organisers on the blind test set (1,000 sentences per direction). Our primary run placed third of four submitted primary systems for EN→TRP, and third of five for TRP→EN.

System	Dir	BLEU	chrF++	TER
Llama-3.2-3B-Instruct (zero-shot)	EN→TRP	0.11	0.76	99.45
Llama-3.2-3B-Instruct (zero-shot)	TRP→EN	0.39	11.93	97.14
KokLLaMA (Primary) [P]	EN→TRP	0.37	16.97	108.06
KokLLaMA (Primary) [P]	TRP→EN	2.04	15.77	136.23

Table 3: KokLLaMA vs. Zero-Shot Baseline on the held-out split ($n = 227$). [P] denotes our primary submission run.

Three observations follow. First, beam search clearly outperforms greedy decoding, by 3.13 BLEU (EN→TRP) and 2.13 BLEU (TRP→EN), which validates the decoding choice described in Section 5.3. Second, EN→TRP is competitive with the field: 5.11 BLEU against a task best of 6.10, and within 1.9 BLEU of the best WMT 2025 system (6.99), despite our system using no parallel training data at all. Third, TRP→EN is *not* competitive, at 3.58 BLEU against a task best of 22.43.

We attribute this asymmetry to the training objective. The instruction corpus teaches the model to *produce* Kokborok, which is precisely what EN→TRP requires, but it provides no signal for producing fluent English conditioned on Kokborok input. In that direction the base model’s strong English prior actively encourages hallucinated paraphrase, as in Example 4 of Table 4. The elevated TER of 200.86 for the TRP→EN primary run is consistent with this reading: the system generates long, fluent, and largely unrelated English. The contrastive run’s lower TER (137.37) at lower BLEU (1.45) confirms that greedy decoding yields shorter but less accurate output rather than better translation.

6.2 Internal Validation Results

Table 3 reports our internal comparison against the zero-shot baseline on the held-out biblical split. A separate comparison of zero-shot against few-shot prompting on a conversational evaluation sample is reported in Appendix C, and our ablation on the effect of post-processing in Appendix B.

KokLLaMA substantially outperforms the zero-shot baseline on chrF++ in both directions, and on TRP→EN BLEU. EN→TRP BLEU remains near-zero for both systems, reflecting fundamental domain divergence. On the TRP→EN task, KokLLaMA achieves a SacreBLEU score of 2.04 and chrF++ of 15.77, compared to 0.39 and 11.93 for the baseline. On EN→TRP, KokLLaMA improves chrF++ from 0.76 to 16.97, providing evidence that conversational fine-tuning successfully transfers

morphological knowledge of Kokborok.

6.3 Reconciling Internal and Official Results

The internal figures in Table 3 are far lower than the official scores in Table 2: 0.37 vs. 5.11 BLEU for EN→TRP, and 2.04 vs. 3.58 for TRP→EN. The system is identical in both cases, so the gap is a property of the evaluation sets rather than of the model.

Two factors explain it. First, domain. Our held-out split is biblical text drawn from the WMT training corpus, whereas the official test set is contemporary news. The instruction corpus is conversational and contemporary, and is therefore much closer in register and vocabulary to news than to scripture. Second, reference style: the biblical references use archaic constructions and dense proper-noun sequences that the model has never observed, which suppresses n -gram overlap disproportionately.

The practical lesson is that a held-out slice of a provided training corpus can be actively misleading as a proxy for shared-task performance when the two differ in domain. Had we relied on our internal numbers alone, we would have concluded that the system had failed almost completely. We report both sets of figures rather than only the more favourable ones.

6.4 Pathological TER Analysis

A key observation in Table 3 is that the zero-shot baseline exhibits the lowest (best) TER scores (99.45 and 97.14). This is an illusion caused by a pathological failure mode of the TER metric. TER is defined as:

$$\text{TER} = \frac{\text{Ins.} + \text{Del.} + \text{Sub.} + \text{Shifts}}{\text{Reference Length}} \times 100$$

Because every extra generated token counts as an insertion (Ins.) edit, TER heavily penalizes models that over-generate or output repetitive tails.

The zero-shot baseline suffers from complete translation failure (BLEU $\approx$ 0.11), meaning it is incapable of outputting Kokborok text. It immediately emits an EOS token or an empty string. Under this empty-generation behavior, the number of required edits is capped at the reference length (all insertions), yielding a TER of exactly $\approx$ 100. This represents a silent failure mode where a non-functional model appears to score well.

In contrast, the fine-tuned KokLLaMA model actively attempts translation but can suffer from over-generation due to low-resource EOS-token

loop failures. We detail how our post-processing pipeline mitigates this and present the corresponding ablation analysis in Appendix B (Table 5).

The official results confirm that this mitigation is only partial. Our EN→TRP primary run scores a TER of 98.67, within the expected range, but the TRP→EN primary run reaches 200.86 under identical post-processing. Because post-processing retains only the first sentence, a TER above 200 implies that the *retained sentence itself* is far longer than the reference, rather than that a repetition tail survived. This points to unconstrained English generation rather than to a decoding loop, and reinforces the argument in Section 5.3 that an explicit `no_repeat_ngram_size` constraint and a tighter `max_new_tokens` budget would have been more appropriate than post-hoc string cleanup.

6.5 Spelling Inconsistency and Oral Nature

Kokborok is historically an oral language, and its writing system using the Roman script has not been fully standardized. As a result, spelling inconsistencies are highly prevalent in parallel datasets. This lack of a standardized orthography heavily penalizes exact-match metrics like SacreBLEU, and helps explain why BLEU scores across the whole Category 2 field remain low (the best EN→TRP system scored 6.10).

Furthermore, our rule-based post-processing cleanup pipeline, while crucial for eliminating repeating loops, can occasionally introduce artifacts. For instance, the script may clip a valid end-clause if it mimics a repetitive tail, leading to a slight drop in BLEU (such as from 0.58 to 0.37 in EN→TRP) even though the readability and usability of the translation are vastly improved. This underscores the need to utilize character-level metrics like chrF++ to capture the true translation quality, which is less sensitive to spelling variations compared to exact-match word BLEU.

6.6 Qualitative Analysis

Table 4 in Appendix A displays qualitative translation samples from KokLLaMA, comparing the raw LLM output against the post-processed prediction. The qualitative analysis reveals a mixed picture of KokLLaMA’s translation capabilities. The model has clearly acquired Kokborok orthography and morphology: its outputs are well-formed and fluent at the surface level, as in Example 2, where it produces grammatical Kokborok with appropriate

case and verb marking. Fluency, however, does not carry semantic fidelity. The Example 2 output diverges substantially from the reference despite being well-formed, and the model also exhibits persistent failure modes. The raw outputs show that the model frequently gets stuck in infinite repetition loops (Example 1), appends numbered word-list annotations (Example 3), or completely hallucinates semantically unrelated content (Example 4, where the prediction describes building a house instead of frogs departing). Although our post-processing cleanup successfully strips many trailing meta-information segments and truncates repeat structures, the model continues to struggle with fluency, loop control, and semantic fidelity in the majority of evaluated cases, including the persistent retention of inline parenthetical glosses (such as ‘(face)’ and ‘(dust)’ in Example 2) compared to standalone note blocks.

Two caveats bound this observation. These examples are drawn from the biblical validation split, which Section 6.1 shows to be the harder of our two evaluation domains, so the failure rate on the official news-domain test set is likely lower. We also did not conduct a systematic human evaluation of the official test outputs; we therefore report the failure rate we observed on the validation split rather than generalising it to the submission as a whole.

7 Conclusion

We presented the SCE-KIIT submission to the WMT 2026 English–Kokborok shared task (Pakray et al., 2026). By fine-tuning Llama-3.2-3B via QLoRA on a conversational instruction corpus containing no parallel translation data, and pairing it with structured prompting and rule-based post-processing, we obtained 5.11 BLEU (EN→TRP) and 3.58 BLEU (TRP→EN) on the official test set, placing third of four and third of five primary systems respectively.

We draw three conclusions. First, cross-task transfer from conversational QA to translation is real but bounded: it suffices to place mid-field in the EN→TRP direction, where the task is to *produce* Kokborok, and is clearly insufficient in the reverse direction, where we trail the best system by 18.9 BLEU. Second, metric interpretation requires care in this regime, since both the baseline’s flattering TER and our own inflated TER are artifacts of generation length rather than of translation

quality. Third, in-domain validation on a slice of the provided training corpus proved a poor predictor of shared-task performance, under-estimating our EN→TRP BLEU by more than an order of magnitude.

Future work should begin with the comparison we were unable to run: fine-tuning on the 2,266 available parallel pairs, alone and in combination with the instruction corpus, to establish whether conversational adaptation complements parallel supervision or merely substitutes for it.

Reproducibility. Fine-tuned model: <https://huggingface.co/agnivamaiti/KokLLaMA-3.2-3B-Instruct>. Corpus: <https://huggingface.co/datasets/agnivamaiti/kokborok-qa>. Code: <https://github.com/AgnivaMaiti/KokborokMT>. The repository implements the prompts of Section 4.2, the post-processing pipeline of Section 4.3, the decoding configuration of Section 5.3 and the deterministic split of Section 3.2, and regenerates Tables 3, 4 and 5. Two artifacts could not be recovered and are documented as such in the repository: the driver script used to train the released adapter, for which we ship a reference implementation reconstructed from Table 1 rather than the original file, and the conversational evaluation sample behind Table 6.

Limitations

No comparison against parallel fine-tuning. The 2,266 official parallel pairs were used only to construct our validation split and few-shot prompts, never for training. We therefore cannot say whether excluding them was beneficial, neutral, or harmful. This is the most important missing experiment in this paper.

No evaluation on the WMT 2025 test set. The organisers released the WMT 2025 test references and encouraged their use. We report published WMT 2025 figures for context, but did not re-run our system on that test set. Our comparison against the 2025 best system is therefore across different test data and should be read as indicative only.

Corpus release mismatch. As noted in Section 3.1, the publicly released instruction corpus is an earlier snapshot of 2,514 distinct pairs, against the 11,428 from which our 10,856 training pairs are drawn, so exact reproduction from public artifacts alone is not currently possible. The conversational

evaluation sample behind Table 6 was likewise not preserved, so that table cannot be regenerated.

Incomplete training records. As noted in Section 4.1, the warmup schedule, hardware configuration, and wall-clock training time were not logged and cannot be reported.

Single run, no variance estimates. All results come from a single fine-tuning run with one random seed. We report no confidence intervals or significance tests, and small BLEU differences on $n = 227$ should not be interpreted as meaningful.

No structured human evaluation. Our qualitative analysis is an informal inspection of validation-split outputs rather than a structured evaluation with native-speaker annotators, adequacy and fluency scales, or inter-annotator agreement. Internal COMET scores were also unavailable because of an unresolved dependency conflict in our evaluation environment; the COMET figures we report come from the organisers.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Fine-tuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Alok Choudhary, Akhil Noulas, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Abteen Ebrahimi, Manuel Mager, and Katharina Kann. 2022. Low-Resource Machine Translation: A Comprehensive Survey. *arXiv preprint arXiv:2203.04561*.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources English-Nyishi pair. *Sādhanā*, 48(4):254.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.
- Graham Neubig and Junjie Hu. 2018. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668.
- Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam

- Jamatia, Koj Sambyo, Ajit Das and Riyanka Manna. 2025. Findings of WMT 2025 shared task on Low-resource Indic Languages Translation. In *Proceedings of the Tenth Workshop on Machine Translation (WMT)*, pages 532–553.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetri-mayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of WMT 2026 shared task on Low-resource Indic Languages Translation. In *Proceedings of the Eleventh Workshop on Machine Translation (WMT)*, co-located with EMNLP 2026, Budapest, Hungary, 28–29 October.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 682–694.
- Maja Popović. 2015. chrF: Character n -gram F-score for Machine Translation Evaluation. In *Proceedings of the Tenth Workshop on Machine Translation*, pages 392–395.
- Maja Popović. 2017. chrF++: words helping character n -grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.
- 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.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Balaraman, Shreya Kundu, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-Resource Languages: A Survey. *ACM Computing Surveys*, 55(11):235:1–38.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

A Qualitative Examples

B Ablation Studies

To analyze the impact of our rule-based post-processing pipeline, we perform an ablation study comparing the translation performance of Kok-LLaMA with and without post-processing. Table 5 presents the detailed SacreBLEU, chrF++, and TER scores on the evaluation split ($n = 227$) for both translation directions (EN→TRP and TRP→EN).

Without post-processing, the model suffers from severe over-generation and repetition loops, which leads to extremely high TER scores of 203.55 (EN→TRP) and 284.84 (TRP→EN). Applying the rule-based post-processing pipeline successfully resolves these repetition loops, reducing the TER scores to 108.06 and 136.23 respectively. In the EN→TRP direction, this truncation occasionally clips valid translation units, resulting in a slight decrease in SacreBLEU (from 0.58 to 0.37) and chrF++ (from 21.07 to 16.97). However, this trade-off is highly favorable as it eliminates pathological repetition loops and produces far more readable and linguistically correct output.

C Conversational Prompting Comparison

Table 6 compares zero-shot against few-shot prompting on a held-out sample of $n = 50$ conversational QA items, scored with the same metrics used throughout the paper. Few-shot prompting achieves a slightly lower BLEU (4.13 vs. 5.36) but a higher chrF++ (34.19 vs. 32.65) and a substantially lower TER (103.60 vs. 130.40). This non-monotonic pattern suggests that in-context examples act mainly as a length and register constraint: they curb the over-generation that inflates TER and improve character-level overlap, while costing a small amount of word-level exact match.

We flag one reproducibility caveat. This 50-item evaluation sample was drawn during development

Source (EN)	The frogs shall depart from you and your houses and your servants and your people; they shall be left only in the Nile.”
Reference	Nwng nini bororok tei nini selengrok yonglani yakni yók bainai. Nil twima-o karwi tei kwrngsa-obo tong rwk glak.”
Raw Output	Chwrai-rok nwngni simi phaiwi nini nok, nini samung tangnai, tei nini borokrokni simi phaiwi kubui khe twima o thahngnai; bohrokni bisingo kubui khe thahngyakhu. Kaisa kaisa jorao bohrokni sakani kubui khe thahngyakhu. Kaisa kaisa jorao bohrokni sakani kubui khe thahngyakhu. Kaisa kaisa
Prediction	Chwrai-rok nwngni simi phaiwi nini nok, nini samung tangnai, tei nini borokrokni simi phaiwi kubui khe twima o thahngnai; bohrokni bisingo kubui khe thahngyakhu.
Source (EN)	In the sweat of your face you shall eat bread till you return to the ground, for out of it you were taken; you are dust, and to dust you shall return.”
Reference	Nono takmani a ha-o kiphilwi thág ya sak nwng khorokni kolomtwi yakungo khiklaiwi tangwi ma chá sinai; nini o hani mang, ha-ono kiphilwi ma thág sinai.”
Raw Output	Nwng chini mukh (face) ni kwanmungni bagwi chahmungni bisingtwi chahmung chahna nangnai, jathatun nwng ha ni simi phainai; nwng kwthar (dust), tei kwthar ni simi phainai. Aboni bagwi nwng chahmungni bisingtwi chahmungni bisingtwi chahmungni bisingtwi chahmungni bisingtwi chahmungni bisingtwi chahm
Prediction	Nwng chini mukh (face) ni kwanmungni bagwi chahmungni bisingtwi chahmung chahna nangnai, jathatun nwng ha ni simi phainai; nwng kwthar (dust), tei kwthar ni simi phainai.
Source (EN)	And Leah conceived again, and she bore Jacob a sixth son.
Reference	Lea sakderiya wngwi tei khoroksa chwla bwsa bá-u kha, abo khorokdok klai-o.
Raw Output	Tei Leah teisa chwngmung wngkha, tei bo Jacob-ni khoroksa haithai (sixth) bwsajla rikhamani. Answer: Tei Leah teisa chwngmung wngkha, tei bo Jacob-ni khoroksa haithai bwsajla rikhamani. 1. 'And' → 'Tei' 2. 'Leah' → 'Leha' 3. 'conceived' → 'chwngmung' 4. 'again' → 'teisa
Prediction	Tei Leah teisa chwngmung wngkha, tei bo Jacob-ni khoroksa haithai (sixth) bwsajla rikhamani.
Source (TRP)	Nwng nini bororok tei nini selengrok yonglani yakni yók bainai. Nil twima-o karwi tei kwrngsa-obo tong rwk glak.”
Reference (EN)	The frogs shall depart from you and your houses and your servants and your people; they shall be left only in the Nile.”
Raw Output	“You will build the house of your ancestors and your relatives with your own hands. Do not go to the river alone and do not go to places that are far away from you.” It is a message about building a house for one’s ancestors and relatives. The last part of the sentence warns against going to places that are far away from one’s home, which is a common practice in many cultures. The message is simple, direct, and practical. It emphasizes the importance of building a house for one’s ancestors and relatives, which is a common practice in many cultures. The last part of the sentence warns against going to places that are
Prediction	You will build the house of your ancestors and your relatives with your own hands.

Table 4: Qualitative translation samples, taken from the first three EN→TRP items and the first TRP→EN item of the held-out validation split. Sources and references are reproduced verbatim from the official parallel corpus; raw outputs are the unmodified model generations, and predictions are the result of applying the pipeline of Section 4.3 to them.

System / Config	Dir	BLEU	chrF++	TER
KokLLaMA (no post-proc)	EN→TRP	0.58	21.07	203.55
KokLLaMA (no post-proc)	TRP→EN	1.22	19.29	284.84
KokLLaMA (with post-proc)	EN→TRP	0.37	16.97	108.06
KokLLaMA (with post-proc)	TRP→EN	2.04	15.77	136.23

Table 5: Ablation results on the evaluation split ($n = 227$), showing BLEU, chrF++, and TER scores.

System	Config	BLEU	chrF++	TER
KokLLaMA	Zero-Shot	5.36	32.65	130.40
KokLLaMA	Few-Shot	4.13	34.19	103.60

and was not preserved alongside the rest of our artifacts. Unlike every other table in this paper, Table 6 therefore cannot be regenerated from the released code and data, and we report it as a development observation rather than as a reproducible result.

Table 6: Results on the conversational QA sample ($n = 50$). Few-shot prompting primes style using training examples.