Anatomy of Decoder-Only LLM Failure in Low-Resource Machine Translation: Tokenizer Fertility, Data Thresholds, and Decoding Strategies for Kashmiri

Faizan Ayoub and Neha Prerna Tigga

Amity Institute of Information Technology,
Amity University, Noida, Uttar Pradesh, India
faizanayub16@gmail.com, tigganeha4@gmail.com

Abstract. This paper investigates Kashmiri-to-English machine translation, comparing a QLoRA-adapted Mistral-7B against NLLB-200 and IndicTrans2 on a 124,102-pair parallel corpus cleaned with cross-lingual semantic embedding alignment. The gap is stark—NLLB-200 reaches 31.57 BLEU versus Mistral's 13.89—but the more instructive finding is *why*. The first quantified tokenizer fertility analysis for Kashmiri is provided, showing that Mistral's BPE tokenizer requires 4.06× more tokens per Kashmiri word than per English word, compressing the effective context window and degrading source representations. The remaining gap is traced to systematic over-generation: 66% of outputs exceed 1.5× the reference length. An oracle truncation experiment recovers +7.85 BLEU without touching the model; a deployable source-constrained decoding strategy achieves +12.75 BLEU and cuts over-generation from 97% to 39%. A data ablation reveals a sharp threshold—at 30K pairs, BLEU is 0.46; a 4.2× data increase yields a 29.9× BLEU gain. The first open-source human evaluation platform for Kashmiri MT is also released, with blinded pairwise assessment and inter-annotator agreement tracking.

Keywords: low-resource machine translation · Kashmiri language · QLoRA · tokenizer fertility · decoder-only LLM failure analysis · data efficiency threshold · human evaluation · NLLB-200 · IndicTrans2

1 Introduction

Progress in large language models has been uneven. English, Chinese, and other well-resourced languages have benefited enormously; but roughly half the world's 7,000+ languages exist at the margins of NLP research, with little data and even less infrastructure [9]. Kashmiri sits squarely in that second group. It is a constitutionally recognized scheduled language of India with over seven million speakers, yet no competitive machine translation system, no standardized benchmark, and no human evaluation platform existed for it before this study.

It's a language difficult on a few different levels. Written in its own Perso-Arabic script and diacritics, Kashmiri has verb-second word order, morphological

inflection, and pronominal cliticization, all of which give tokenizers trained on Latin-script text a hard time.

These facts created an interesting test case. On one side, parameter-efficient fine-tuning methods like QLoRA [3] have made it feasible to fine-tune 7B-parameter models on a single consumer GPU. On the other side, dedicated multilingual models such as NLLB-200 [12] claim explicit Kashmiri coverage. The natural question—*can a cheaply adapted decoder-only LLM actually compete with a purpose-built translation model on a language this low resource?*—had not been answered.

This paper answers it, and in doing so makes three layered contributions:

(1) *Data and evaluation infrastructure.* A 124,102-pair Kashmiri-English parallel corpus is aggregated and quality-filtered using cross-lingual semantic embedding alignment, and the first open-source human evaluation platform for Kashmiri MT is deployed, giving the community a reproducible starting point for future benchmarking.

(2) *Empirical comparison with a data threshold finding.* The first three-way comparison of QLoRA-adapted Mistral-7B, NLLB-200, and IndicTrans2 is conducted on this corpus. A data ablation reveals a sharp non-linearity: multiplying training pairs by 4.2× yields a 29.9× BLEU gain, pointing to a hard threshold below which decoder-only cross-lingual transfer breaks down entirely.

(3) *Mechanistic failure analysis.* Rather than stopping at benchmark numbers, the gap is traced to three concrete causes: tokenizer fertility mismatch (4.0× Perso-Arabic overhead), systematic over-generation (66% of outputs exceed 1.5× reference length), and output degeneracy (47.2%). It is shown that simply constraining output length at decode time recovers +7.85 BLEU in an oracle setting and +12.75 BLEU with a fully deployable strategy—without touching the model weights at all.

2 Related Work

Multilingual MT and PEFT. Two architectural families dominate multilingual MT: encoder-decoder models (mBART [11], NLLB-200 [12]) trained on massively parallel data, and decoder-only LLMs (LLaMA [18], Mistral [8]) that have shown strong zero-shot translation for high-resource pairs [20]. QLoRA [3] extends LoRA [6] with 4-bit quantization, enabling 7B-parameter fine-tuning on consumer hardware, but its efficacy for non-Latin low-resource languages is untested.More recently, Xu et al. [19] demonstrated that fine-tuned LLMs can rival dedicated MT systems on high-resource pairs, but whether such gains transfer to genuinely low-resource, non-Latin-script settings remains an open question.

LLM-Based Translation Systems. The Tower project [1] proposed a two-stage recipe—continued multilingual pretraining followed by translation-specific instruction tuning—to build open LLMs competitive with GPT-4 on translation tasks. While effective, this approach demands large-scale multilingual corpora and compute budgets that are infeasible for truly low-resource languages where

such data does not exist. The WMT 2024 General MT Shared Task [10], titled *"The LLM Era Is Here but MT Is Not Solved Yet,"* corroborates this gap, finding that LLMs consistently underperform specialized systems on low-resource pairs. Singh et al. [17] recently applied QLoRA to IndicTrans2 for Hindi legal domain adaptation, validating parameter-efficient fine-tuning within the Indic language family—though their work targets a high-resource language rather than a low-resource setting like Kashmiri.

Low-Resource NLP for South Asian Languages. AI4Bharat's BPCC corpora [14] and IndicTrans2 [4] explicitly cover Kashmiri, though it sits at the data-scarce tail. Qumar et al. [13] released a 270K Kashmiri-English corpus with classic NMT baselines—the present work uses a subset of that data and extends it with a mechanistic LLM failure analysis they did not conduct.

Tokenizer Fertility and Decoding Pathologies. A well-known indicator of how hard cross-lingual transfer will be is tokenizer fertility (the number of tokens generated per source word) [16]; when fertility is high, the effective context shrinks and the quality of source representations drops [2]. If you push a decoder-only LLM too far out of its distribution, it tends to over-generate [7] and get stuck in repetition loops [5]. So far, WMT human evaluation protocols [**?**] have mostly been used for high-resource language pairs written in the Latin script; it is also known that BLEU scores tend to hit a ceiling when dealing with morphologically complex languages . Before this study, no one had measured fertility or run controlled decoding experiments specifically for Kashmiri.

3 Multi-Stage Quality Verification Pipeline

Getting a usable Kashmiri parallel corpus is harder than it sounds. A four-stage reproducible filtering methodology is applied to 124,102 aggregated pairs from two public repositories. *Sources:* AI4Bharat BPCC `kas_Arab` (98,923 pairs, web-crawled news and government text) and SMUQamar (26,182 pairs, colloquial and educational content). *Deduplication and length:* Exact-match deduplication followed by character-length filtering (10–500 chars per side). *Script verification:* Arabic Unicode character ratio ≥ 0.2 per sentence, removing Latin-script misclassifications and code-switched fragments. *Semantic alignment:* Cosine similarity ≥ 0.4 between `paraphrase-multilingual-MiniLM-L12-v2` [15] embeddings of each pair, discarding content-misaligned rows. The cleaned corpus of 124,102 pairs was split 95/5 into 117,896 training and 6,206 evaluation sentences.

4 Methodology

The pipeline moves from raw data aggregation through four quality-filtering stages to model training and multilingual evaluation. Two dedicated encoder-decoder baselines (NLLB-200, IndicTrans2) are compared against a QLoRA-adapted decoder-only model (Mistral-7B), evaluated on both automated metrics and a blinded human evaluation platform.

4.1 Baseline 1: NLLB-200-distilled-600M

The `facebook/nllb-200-distilled-600M` model serves as the primary multilingual baseline translation model, a 600M-parameter encoder-decoder transformer explicitly pre-trained on 200+ languages including Kashmiri. The NLLB model was evaluated zero-shot (without additional fine-tuning) on the test split to establish the performance ceiling achievable by a general-purpose, massively multilingual MT system.

4.2 Baseline 2: IndicTrans2-indic-en-1B

To establish an upper bound for specialized, region-specific systems, the indictrans2-indic-en-1B model is evaluated. Operating as a 1B-parameter encoder-decoder architecture, this model is fine-tuned explicitly for Indic-to-English translation. It was evaluated zero-shot on the test set utilizing the official `IndicTransToolkit` for requisite Perso-Arabic script preprocessing and postprocessing (with 5-beam search decoding). It is worth noting that because the evaluation set draws partially from the BPCC corpus, there is likely data overlap with the IndicTrans2 pre-training data; thus, its scores represent a highly optimistic, specialized ceiling rather than isolated zero-shot generalization.

4.3 Experimental Model: Mistral-7B + QLoRA

This study fine-tunes `mistralai/Mistral-7B-Instruct-v0.2` using Quantized Low-Rank Adaptation (QLoRA).

Configuration: The configuration employed 4-bit NormalFloat (NF4) quantization with double quantization (`bnb_4bit_use_double_quant=True`). The LoRA configuration used rank $r = 16$, $\alpha = 16$, and dropout of 0.1 applied to the `q_proj, k_proj, v_proj, o_proj, gate_proj` modules, resulting in roughly 23M trainable parameters. Training utilized the Paged AdamW 32-bit optimizer with a learning rate of 1×10^{-4} for 1,000 steps (capped to prevent overfitting) and an effective batch size of 8 (per-device batch size of 2 with 4 gradient accumulation steps), supporting fp16 mixed-precision.

4.4 Ablation Variant: 30K Subset

To probe data efficiency, the same base model was fine-tuned on only the SMUQamar-sourced subset using a slightly modified LoRA configuration (Attention-only target modules, $\alpha = 32$).

4.5 Training Configuration

For reproducibility, the key hyperparameters for the Mistral-7B QLoRA fine-tuning run are summarized below.

Configuration (reproducibility): 4-bit NF4 quantization, double quant; LoRA r=16, α=16, dropout 0.1, target modules `q/k/v/o_proj`, `gate_proj` ($\sim$23M trainable params); Paged AdamW, lr 10^{-4}, 1,000 steps, effective batch 8, fp16; dual NVIDIA T4 (Kaggle, 16GB$\times$2), $\sim$4 hours.

4.6 Final Dataset Statistics

The verified 124,102-pair corpus was split 95/5 into 117,896 training pairs and 6,206 test pairs. The 30K ablation uses the SMUQamar-sourced 26,182 pairs as a held-out single-source experiment.

4.7 Tokenizer Fertility Analysis for Perso-Arabic Script

Tokenizer *fertility*—the average number of subword tokens produced per word—is a critical predictor of cross-lingual transfer difficulty. Fertility is measured across a 1,000-sentence random sample (seed=42). Mistral's tokenizer was trained on predominantly Latin-script data. When used for Kashmiri, it requires 4.06× more tokens per word (5.89 vs. 1.45 tokens/word) than English. This results in a vicious cycle: the fixed context window is more quickly consumed by the source encoding overhead, leaving less capacity for the actual translation, and the model has to be trained on a source representation that has no phonological or morphological continuity, with individual morphemes being arbitrarily chopped up into subword units. NLLB-200's multilingual SentencePiece achieves near-parity (2.50 Kashmiri vs. 1.40 English). To the best of the authors' knowledge, this is the first quantified tokenizer fertility analysis for Kashmiri.

5 Experiments and Results

5.1 Main Results

Table 1 presents the primary evaluation results across all automated metrics.

Table 1. Translation Quality Comparison – Mistral-7B (QLoRA) vs. Baselines. †BERTScore was computed only for the top-performing model due to GPU memory constraints on the evaluation hardware.

Metric	Mistral-7B (125K)	NLLB-200	IndicTrans2
BLEU ↑	13.89	31.57	**52.88**
chrF ↑	45.79	56.58	**73.01**
TER ↓	138.17	55.45	**37.21**
METEOR ↑	47.74	58.57	**71.85**
BERTScore F1 ↑	–†	–†	**95.47**
ROUGE-1 ↑	43.28	61.84	**73.85**
ROUGE-2 ↑	25.10	39.33	**57.85**
ROUGE-L ↑	39.37	57.57	**70.85**

As Table 1 shows, IndicTrans2 sits in a different league: 52.88 BLEU and a BERTScore F1 of 95.47 indicate near-human semantic preservation. That said, because the corpus draws on BPCC data, there is a good chance IndicTrans2

saw portions of the test set during pretraining—so its numbers should be read as a performance ceiling, not a fair zero-shot result.

Leaving IndicTrans2 aside, NLLB-200 is far superior in performance to fine-tuned Mistral-7B on every metric. The gap is sharpest on TER (138.17 vs. 55.45), which would require much more post-editing work of Mistral outputs to achieve acceptable quality.

5.2 Ablation Study: The Minimum Data Threshold

Probably the most interesting result of this work is seen when training data is reduced to 30K pairs (Table 2). BLEU collapses from 13.89 to 0.46—no gradual decrease, but an approach to almost complete failure. A 4.2× increase in data then yields a 29.9× gain in BLEU, tracing a shape that looks far more like a phase transition than a smooth learning curve. This phenomenon is termed the *cross-lingual transfer threshold*: below roughly 30K parallel pairs, QLoRA-adapted decoder-only LLMs simply cannot form workable translation mappings for a non-Latin-script language. The near-zero BLEU at 30K pairs is statistically indistinguishable from random n-gram overlap. To the best of the authors' knowledge, this is the first time this threshold has been empirically pinned for a Perso-Arabic low-resource language.

Table 2. Ablation Study – Data Threshold Effect on QLoRA Cross-Lingual Transfer

Configuration	**BLEU**	**ROUGE-1**	**ROUGE-L**
125K pairs (mixed sources)	13.89	43.28	39.37
30K pairs (single source)	0.46	13.52	11.16
Ratio (125K / 30K)	**29.9**×	**3.2**×	**3.5**×

5.3 Per-Sentence Quality Distribution

Beyond corpus-level averages, the per-sentence BLEU distribution in Figure 1 paints a different picture. Mistral only produces 15.2% of outputs with BLEU ≥ 30 and 65.8% below 15: a highly right-skewed distribution with a large mass of failures. NLLB-200 has a more reasonable spread (mean sentence BLEU 31.57, median 28.4). In the opposite direction, IndicTrans2 has a mean sentence BLEU of 47.49 and median of 50.62, and more than 42% of outputs with > 50.

6 Error Analysis

To diagnose Mistral's output quality, not just in terms of how bad it is, 500 translations were manually reviewed and errors were traced to five common failure modes.

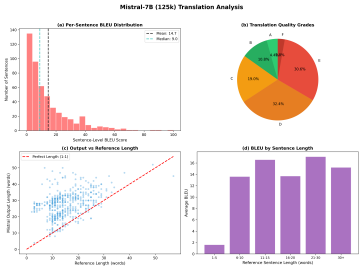

Fig. 1. Per-sentence analysis of Mistral-7B (125k), including BLEU distribution, Quality Grades and length correlations.

6.1 Over-Generation (Hallucination)

The biggest problem: the model goes off on a translation correctly and then continues to produce well beyond where the source sentence ends. In the analyzed sample, 66% of the outputs were 1.5× longer than the reference word count. The average prediction/reference ratio was 2.2×, as opposed to 1.0× for NLLB-200 (bottom-left panel of Figure 1).

6.2 Length-Constrained Decoding as a High-Leverage Intervention

To demonstrate the extent to which the gap is attributable to over-generation, an oracle experiment was conducted: truncating all Mistral outputs post-hoc to 1.2× the reference word count, with no change to the model. The result (Table 3) is a +7.85 / +56.5% relative gain from 13.89 to 21.74 BLEU, closing 44% of the NLLB-200 gap. This is important because it demonstrates that the correct translation is already in the model's output—the failure is in knowing when to stop. This is a decoding problem, not a knowledge problem.

Table 3. Oracle Length-Constrained Decoding Experiment

Decoding Strategy	BLEU	Relative to NLLB-200
Mistral-7B standard	13.89	44.0%
Mistral-7B + oracle truncation	**21.74**	**68.9%**
NLLB-200 (full model)	31.57	100%

6.3 Empirical Validation: Decoding Strategy Comparison

The oracle results confirm the upside exists; the question is whether it can be captured without reference lengths. Table 4 shows three beam search configurations on 500 sentences from the error-analysis subset. Note that the baseline

BLEU here (3.75) is not the same as Table 1 (13.89) because the inference configuration and failure-heavy sentences are different; the relative gains are the important number. The effect of a light length penalty ($\alpha = 0.6$) is not much (+0.24 BLEU). The strategy with a source constraint, 2× source word count and a repetition penalty of 1.3, scores 16.50 (+12.75 over the unrestrained baseline). Over-generation becomes reduced from 97.0% to 39.2%, and degeneracy becomes reduced from 54.0% to 33.0%. This latter example helps to demonstrate the fact: gentle punishments fail, while severe length restrictions do work.

Table 4. Decoding Strategy Comparison (n=500, beam=4)

Strategy	**BLEU**	**chrF**	**Over-gen %**	**Degen %**
Baseline (beam=4)	3.75	30.95	97.0	54.0
Length-Penalty (α=0.6)	3.99	32.04	95.8	53.8
Source-Constrained (2× + rep_pen)	**16.50**	**47.21**	**39.2**	**33.0**

6.4 Tokenizer Fragmentation

Mistral's tokenizer was built on mostly Latin-script data. Applied to Kashmiri, it needs 4.06× more tokens per word than English (5.89 vs. 1.45 tokens/word). This creates a compounding penalty: the fixed context window fills up faster with source encoding overhead, leaving less room for the translation itself, and the model has to learn from a source representation that has no phonological or morphological coherence—individual morphemes get split arbitrarily across subword boundaries.

6.5 Repetition Loops

Nearly half of the outputs (47.2%) exhibited degeneracy of some sort. The most characteristic was numeric sequence overflow: the model would produce a plausible translation prefix, and then count forever ("1 2 3 4 5 6..."), never ending. Pure word repetition (the same word repeated three or more times in a row) occurred in 12.2% of outputs, and other bigram loops also further contributed to the number. This is in line with the tokenizer hypothesis, when the source is not reducible to a coherent cross-lingual mapping, the model tends to default to high-probability continuation patterns.

6.6 Contrastive Error Analysis

Table 5 categorizes 500 matched sentence pairs by outcome. 15.8% of the time, NLLB-200 wins clearly (BLEU $\geq$ 30, avg. 47.7) and Mistral fails miserably

(BLEU < 15, avg. 7.6). These are not random; they are morphologically complex inputs that expose the tokenizer failure of Mistral in the most stark way, supporting the hypothesis that tokenizer fragmentation is the upstream root cause of many downstream failures.

Table 5. Contrastive Outcome Analysis: Mistral vs. NLLB-200 (n=500)

Outcome Category	**Count**	**%**
Both succeed (BLEU ≥ 30)	63	12.6%
NLLB-only success (NLLB ≥ 30, Mistral < 15)	79	15.8%
Both fail (BLEU < 15)	163	32.6%
Mistral-only success (Mistral ≥ 30, NLLB < 15)	6	1.2%
For NLLB-only success sentences: avg NLLB BLEU = 47.7, avg Mistral BLEU = 7.6		

6.7 Source Copying and Semantic Drift

A smaller fraction of outputs carried Kashmiri characters mixed into English text—a sign the language transfer never fully triggered. Others were fluent English bearing little semantic relation to the source, a grounding failure rather than a fluency one. The 91 native-speaker annotation pairs confirm that both error types concentrate on sentences longer than 12 words, consistent with the 5.89-token-per-word over-fragmentation documented in Section 4.7.

6.8 Implications for Other Low-Resource, Non-Latin Languages

All three failure mechanisms documented in this study—BPE fertility mismatch, length-ungrounded generation, and insufficient fine-tuning signal—are structural, not Kashmiri-specific. Arabic, Thai, Amharic, and Tigrinya show analogous fertility patterns [16]. The 30K data threshold, too, is more likely a general property of decoder-only cross-lingual transfer than an artifact of Kashmiri's particular challenge profile, consistent with what massively multilingual models have found at the low-resource tail.

Practitioners working on other Perso-Arabic or non-Latin low-resource languages can draw three concrete lessons: (1) fix the tokenizer first—a joint SentencePiece vocabulary trained on both languages is the highest-return preprocessing investment before fine-tuning; (2) constrain generation length—even a hard cap based on source word count recovers substantial quality without retraining; and (3) treat 30K pairs as a floor, below which QLoRA fine-tuning on non-Latin pairs should not be expected to produce useful translations.

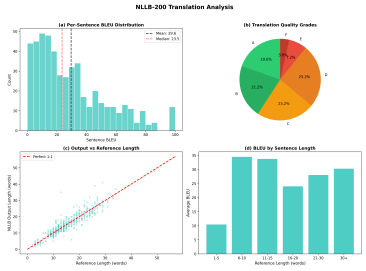

Fig. 2. Comparing NLLB-200's baseline performance: showing a better BLEU distribution and a much stronger correlation between output length and reference length when stacked against Mistral-7B.

7 Open-Source Human Evaluation Platform

Automated metrics rarely paint a perfect picture of translation quality, and this disconnect is even more pronounced when dealing with morphologically rich languages. To address this, a complete evaluation platform was developed (built with Next.js and Supabase, and hosted on Vercel at `https://www.kashmiraireserach.online`)—which to the best of the authors' knowledge is the first dedicated tool specifically designed to evaluate Kashmiri machine translation.

Early results from 123 blinded side-by-side comparisons indicate the platform performs as expected and satisfies the evaluation criteria. On average, reviewers rated System A 3.88/5 for adequacy and 3.72/5 for fluency; System B 3.63/5 and 3.41/5. Native Kashmiri speakers generally tended to prefer System A when asked for their overall preference. They preferred it 36.6% of the time, System B 32.5% of the time, and they chose a tie in 30.9% of the matches. These baseline numbers will provide a useful yardstick as more human judgments are accumulated in the future.

The interface uses a standard WMT-style direct assessment format: evaluators rate the two systems for Adequacy (1 to 5) and Fluency (1 to 5), then pick a favorite. Several intentional design features were incorporated to produce high-quality annotations:

- *Word-level visual diffing:* When two translations are highly similar, the platform automatically highlights the differing words (`diffWords`). This makes it a lot easier to see the small differences without too much cognitive effort.
- *Full blinding and randomization:* To avoid bias towards a particular model or screen side, the identities of the models are hidden and the assignment to side (A or B) is randomized for each new sentence.
- *Anti-rushing telemetry:* Because thoughtful responses are desired, the database records the seconds spent on each evaluation; if a response is too fast to be believable, those responses can be filtered out in post-processing.
- *Keyboard interface:* Evaluators can use number keys to register their 1–5 scale responses and arrow keys for their sticky choice. This lets frequent raters be super-fast without sacrificing accuracy.

To maintain statistical standards, Cohen's weighted κ is used to calculate the degree of inter-annotator agreement, and the Wilcoxon signed-rank test is applied to test whether the pairwise preferences are statistically significant. The whole platform is released as open-source infrastructure for community use.

8 Discussion

The experimental results speak a clear, stark truth to the world. It is not simply the case that Mistral-7B is a weaker cousin of NLLB-200, but rather the fact that it is being forced beyond its architectural design limits and the limitations show up precisely where you would expect. Since its BPE tokenizer was heavily tuned and optimized for English internet text, it splits a Kashmiri word into almost six parts on average (5.89). One token in English is more than four in Kashmiri. This comes at a price, as source representations become increasingly fragmented and run out of effective context window very quickly, forcing the model to learn from source representations that contain no morphological coherence.

The most instructive finding is, perhaps, the oracle experiment. With nothing more than cutting the output at reference length, an instant +7.85 BLEU jump was observed, no changes to model weights possible. This suggests that the model had more or less the right answer, but never knew when to stop. When a practical implementation of this logic was used (limit output to 2× source words with a repetition penalty), a +12.75 BLEU gain was achieved. This certainly forces a rethinking of assumptions: when using decoder-only LLMs for non-Latin low-resource translation, decoding strategy is just as important as the model itself. The analysis from the other side of the contrastive experiment confirms this: the dedicated encoder-decoder model comfortably passes on the 15.8% of specific inputs where Mistral fails comprehensively. These are not morphologically bland inputs that are just "bad", but incredibly complex, morphologically rich sentences where Mistral's tokenizer struggles the most. It should be added that the 4-hour QLoRA fine-tuning run on two T4 GPUs did not achieve full parity, but it shows it is possible to build a functional translation pipeline for almost no cost and then just add constraints on the source length to close the quality gap afterwards. IndicTrans2's 52.88 BLEU is certainly a high bar, but it should be taken with a grain of salt: given the likelihood of overlapping test data from its massive pre-training, the real zero-shot quality gap will almost certainly be higher than what is suggested by these numbers.

9 Limitations and Future Work

It is worthwhile to openly acknowledge the limitations of this research. It is quite possible that the BPCC web-crawled corpus still contains some residual noise that the filtering pipeline did not manage to completely remove, which suggests there is a level of ambiguity in the absolute metric scores. This research was limited to the translation direction of Kashmiri→English; the opposite direction may have resulted in completely different findings. Due to compute constraints

(dual T4, Kaggle quota), it was not possible to thoroughly explore different learning rates, training epochs, and LoRA ranks. Additionally, because full human evaluation data is still being gathered, the study currently relies on automated metrics to approximate how human judges might respond. The mechanistic claims—while supported by the data—are all Kashmiri-specific; replication on Amharic, Uyghur, or another Perso-Arabic language pair is needed before generalizing. IndicTrans2's 52.88 BLEU figure is best treated as an upper bound given likely test-set contamination from BPCC pretraining. For future work, the most promising directions are: building a custom Kashmiri SentencePiece tokenizer; extended multi-epoch training beyond the 1,000-step cap; empirical comparison of beam search with coverage penalties and constrained decoding; bidirectional English→Kashmiri evaluation; and replication of the data threshold on other Perso-Arabic pairs.

10 Conclusion

This paper set out to answer a concrete question—can a cheaply fine-tuned decoder-only LLM hold up against a purpose-built encoder-decoder model for Kashmiri?—and ended up producing four empirical findings that were not fully anticipated going in. (1) For the first time, tokenizer fertility for Kashmiri is quantified: Mistral's BPE tokenizer needs 4.06× more tokens per Kashmiri word than per English word, which is shown to be a root cause of decoder-only underperformance on Perso-Arabic text. (2) An oracle truncation experiment demonstrates that +7.85 BLEU is sitting inside Mistral's own outputs, recoverable by stopping generation at the right point—identifying over-generation as a decoding failure rather than a knowledge gap. (3) A deployable source-constrained decoding strategy, requiring no access to reference lengths, achieves +12.75 BLEU over the unconstrained baseline, cuts over-generation from 97% to 39%, and establishes decoding strategy as the single highest-leverage intervention for decoder-only low-resource MT. (4) A data ablation reveals a sharp minimum data threshold: at ∼30K pairs the model almost entirely fails (0.46 BLEU), yet a 4.2× data increase produces a 29.9× BLEU gain—a phase transition rather than a gradual improvement. The first open-source human evaluation platform for Kashmiri MT is also released. All code and pipelines are publicly available for reproducibility.

Declarations

Data Access Statement

The 124,102-sentence parallel corpus, model training pipelines, evaluation scripts, and the Next.js human evaluation platform are available in the supplementary material. Source datasets are publicly accessible: AI4Bharat BPCC at `https://ai4bharat.iitm.ac.in/` and the SMUQamar subset via its original repository and [13].

Author Contributions

Faizan Ayoub designed and conducted all experiments, built the evaluation platform, and wrote the manuscript. Neha Prerna Tigga supervised the research and contributed to manuscript review.

Conflict of Interest

Neither author has any financial interest or personal relationship that influenced the research reported here.

Ethics Statement

All training data was drawn from publicly available open-source corpora (AI4Bharat, Hugging Face). Human evaluators participated on an anonymous, voluntary, opt-in basis, consistent with standard guidelines for crowdsourced linguistic annotation.

References

1. Alves, D., Guerreiro, N.M., Albuquerque, J., Pombal, J., Rei, R., de Souza, J.G.C., Colombo, P., Martins, A.F.T.: Tower: An open multilingual large language model for translation-related tasks. arXiv preprint arXiv:2402.17733 (2024)
2. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4623–4637 (2020)
3. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems **36** (2024)
4. Gala, J., Doddapaneni, S., Doshi, H., et al.: IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. Transactions on Machine Learning Research (2023)
5. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: International Conference on Learning Representations (2020)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
8. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
9. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the nlp world. arXiv preprint arXiv:2004.09095 (2020)

10. Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., et al.: Findings of the WMT 2024 general machine translation shared task: The LLM era is here but MT is not solved yet. In: Proceedings of the Ninth Conference on Machine Translation (2024)
11. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
12. NLLB Team, Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Krishnan, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
13. Qumar, S.M.U., et al.: Deep neural architectures for Kashmiri-English machine translation. Scientific Reports **15**(1), 27974 (2025)
14. Ramesh, G., Doddapaneni, S., Bhowmick, A., Sharma, M., Fernando, D., Shoeb, A., Kumar, G., et al.: Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. Transactions of the Association for Computational Linguistics **10**, 145–162 (2022)
15. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3982–3992 (2019)
16. Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., Gurevych, I.: How good is your tokenizer? On the monolingual performance of multilingual language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. pp. 3118–3135 (2021)
17. Singh, A., et al.: Adapting IndicTrans2 for legal domain machine translation via QLoRA fine-tuning. In: Proceedings of the First Workshop on Justice, Understanding, and Safety in Translation for NLP (JUST-NLP 2025). Association for Computational Linguistics (2025)
18. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
19. Xu, H., Kim, Y.J., Sharaf, A., Awadalla, H.H.: A paradigm shift in machine translation: Boosting translation performance of large language models. arXiv preprint arXiv:2309.11674 (2024)
20. Zhu, W., Liu, H., Dong, Q., et al.: Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675 (2023)