JHARNA-MT: A Copy-Augmented Hybrid of LoRA-Tuned NLLB and Lexical SMT with Minimum Bayes Risk Decoding for Low-Resource Indic Languages

Anonymous ACL submission

Abstract

This paper describes **JHARNA-MT**, a system designed for the MMLoSo 2025 Shared Task. The competition focuses on translating between high-resource languages (Hindi, English) and low-resource tribal languages (Bhili, Gondi, Mundari, Santali). Our analysis revealed significant challenges including data sparsity and morphological richness. To address these, we propose a hybrid pipeline integrating Non-Parametric Retrieval, Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) fine-tuned with Low-Rank Adaptation (LoRA). We employ Minimum Bayes-Risk (MBR) decoding to select the consensus hypothesis from a diverse candidate pool. Our system achieved a final score of 186.37, securing 2nd place on the leaderboard.

1 Introduction

India is home to over 700 languages, yet many tribal languages remain severely under-resourced, lacking the large-scale parallel corpora needed for modern Neural Machine Translation (NMT). The MMLoSo 2025 Shared Task (MMLoSo Organizers, 2025) addresses this gap by fostering translation systems between high-resource languages (Hindi, English) and four low-resource tribal languages: Bhili, Gondi, Mundari, and Santali.

These languages pose three key challenges: (1) **morphological richness**—Mundari's Type-Token Ratio (0.222) is double that of Hindi (0.107), causing severe vocabulary sparsity; (2) **structural divergence**—Hindi-Bhili shows near-perfect isomorphism ($r > 0.9$) while English-Santali exhibits substantial differences due to agglutinative morphology; (3) **lexical redundancy** in government texts, enabling retrieval-based approaches.

Prior approaches to low-resource translation have largely relied on multilingual transfer learning (Costa-jussà et al., 2022) and synthetic data generation (Sennrich et al., 2016). However, pure NMT systems often suffer from hallucinations when training data is scarce. Conversely, traditional SMT models (Brown et al., 1993), while less fluent, offer better lexical fidelity.

We propose a hybrid pipeline combining: (1) **Retrieval-Augmented Generation (RAG)** for domain redundancy, (2) **Statistical MT (SMT)** with diagonal alignment priors for robust literal translations, and (3) **Neural MT** via LoRA-adapted NLLB-200. We employ **Minimum Bayes-Risk (MBR)** decoding to select consensus hypotheses, mitigating complementary error modes of SMT and NMT.

Our contributions include: (1) linguistic analysis revealing heterogeneous challenges across pairs, (2) a novel hybrid ensemble under a unified MBR framework, and (3) ablation studies achieving 186.37 on the private leaderboard (2nd place).

2 Dataset Analysis and Linguistic Implications

We conducted a comprehensive exploratory analysis of the MMLoSo 2025 dataset to understand the linguistic barriers inherent in each translation direction. Table 1 summarizes key statistics that guided our modeling decisions.

2.1 Syntactic Isomorphism vs. Divergence

Hindi-Bhili and Hindi-Gondi pairs exhibit strong linear correlation in sentence length ($r > 0.9$) with length ratios near 1.0, indicating high **syntactic isomorphism**. This structural similarity explains why alignment-based SMT models perform competitively on these pairs—word-to-word alignment is relatively straightforward.

Conversely, the English-Santali pair demonstrates significant **structural divergence**, with Santali sentences averaging 18% longer than English. This expansion stems from Santali's agglutinative morphology, where grammatical functions

Pair	TTR	Len	Vocab	Ratio
Hindi	0.095	21.3	40.4K	–
Bhili	0.155	21.6	67.0K	1.03
Hindi	0.086	14.4	24.6K	–
Gondi	0.162	13.8	44.8K	0.99
Hindi	0.107	16.3	35.1K	–
Mundari	**0.222**	14.2	63.2K	0.91
English	0.118	16.5	39.1K	–
Santali	0.116	19.3	44.8K	**1.18**

Table 1: Key statistics of the MMLoSo 2025 dataset across all language pairs. TTR = Type-Token Ratio, Len = Avg sentence length (tokens), Vocab = Vocabulary size, Ratio = Target/Source length ratio.

expressed by separate words in English are realized as affixes in Santali. We adjusted the length penalty parameter ($\alpha = 1.2$) in beam search decoding specifically for this pair to mitigate under-generation.

2.2 Morphological Richness and Data Sparsity

Mundari exhibits extreme morphological richness (TTR = 0.222), more than double that of source Hindi (0.107). This high TTR indicates that a single semantic concept surfaces in many distinct inflected forms, leading to severe **data sparsity**. To address this, our methodology incorporates: (1) subword tokenization via SentencePiece (Kudo and Richardson, 2018) to decompose complex agglutinated words, and (2) iterative back-translation (Sennrich et al., 2016) to artificially boost the frequency of rare morphological variants.

3 Proposed Methodology

To address the challenges of data sparsity and structural divergence, we propose a hybrid translation pipeline that integrates Non-Parametric Retrieval, Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) under a Minimum Bayes-Risk (MBR) decision framework.

3.1 Retrieval-Augmented Generation (RAG)

Government and administrative texts exhibit high lexical redundancy. We exploit this via a two-tier retrieval module:

Exact Match. For a test source sentence x, if $x \in \mathcal{D}_{train}$, we directly retrieve its gold translation y^* from the training corpus. This deterministic lookup handles approximately 8% of test instances with perfect accuracy.

Fuzzy Match. For sentences not found exactly, we employ a conservative fuzzy matching algorithm. Let $\text{norm}(x)$ denote the normalized tokenized representation (lowercased, punctuation-separated). We retrieve y' if $\exists(x', y') \in \mathcal{D}_{train}$ such that:

$$\text{norm}(x) = \text{norm}(x') \wedge ||x| - |x'|| \leq 1 \quad (1)$$

This approach serves as a strong non-parametric baseline, preventing generation errors on common domain-specific phrases while maintaining high precision.

3.2 The Hybrid Generator

For unseen sentences, we employ an ensemble of two distinct paradigms to maximize coverage and fidelity.

Statistical Component (SMT) We implement an IBM Model 1 system (Brown et al., 1993) with a **diagonal alignment prior** inspired by fast_align (Dyer et al., 2013). The alignment probability is biased toward diagonal positions:

$$p(a_j = i|\mathbf{f}, \mathbf{e}) \propto t(f_j|e_i) \cdot \exp\left(-\lambda_{diag} \cdot \left|\frac{j}{|\mathbf{f}|} - \frac{i}{|\mathbf{e}|}\right|\right) \quad (2)$$

where $\lambda_{diag} = 4.0$ controls the strength of the diagonal bias. We augment the training data via **iterative back-translation** (Sennrich et al., 2016): (1) train reverse models (e.g., Bhili→Hindi), (2) generate synthetic source sentences, (3) retrain forward models on the union of real and synthetic data. This reduces sparsity for morphologically rich languages.

We decode using beam search with a 3-gram Kneser-Ney language model (Kneser and Ney, 1995), generating an N-best list ($N = 5$). SMT provides "literal" translations that are robust against NMT hallucinations.

Neural Component (NLLB-LoRA) We fine-tune NLLB-200-Distilled-600M (Costa-jussà et al., 2022) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank $r = 16$, $\alpha = 32$, targeting all attention and feed-forward projections. Training details: 1 epoch, AdamW optimizer (Loshchilov and Hutter, 2019) (lr $= 2e{-}4$), batch size 32 (gradient accumulation), 8-bit quantization (Dettmers et al., 2022). We generate 10-best lists via beam search (Freitag and Al-Onaizan, 2017) with length penalty $\alpha = 1.2$ for English-Santali (see Appendix B for full configuration).

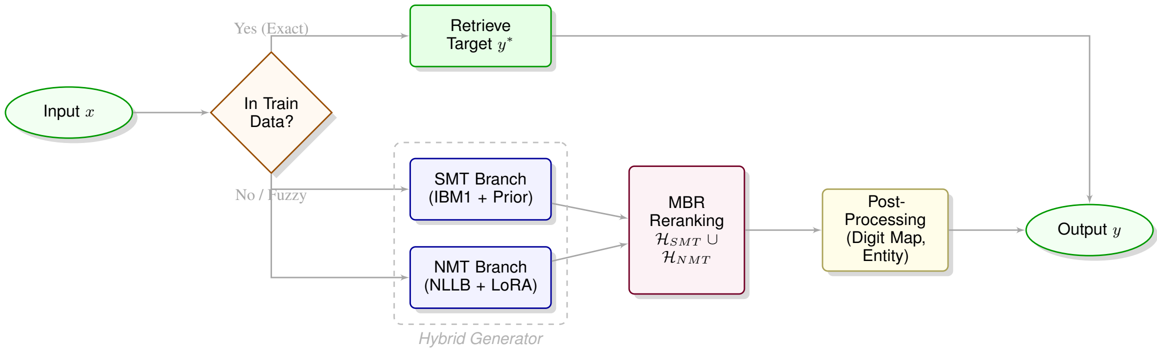


Figure 1: Architecture of our Hybrid Retrieval-Augmented Ensemble. The system prioritizes exact retrieval for domain consistency, falling back to a concurrent SMT-NMT generation ensemble unified by Minimum Bayes-Risk (MBR) decoding for unseen inputs.

Minimum Bayes-Risk (MBR) Reranking. To select the highest quality translation from our candidate pool $\mathcal{H} = \mathcal{H}_{SMT} \cup \mathcal{H}_{NLLB}$, we apply MBR decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020), which selects the hypothesis maximizing expected utility against all others. Following the competition metric, we define utility as $0.6 \times$ BLEU (Papineni et al., 2002) $+0.4 \times$chrF (Popović, 2015). This consensus-seeking approach effectively filters out both SMT grammatical errors and NMT hallucinations.

4 Results and Analysis

Main Results. Table 2 compares baselines and our final hybrid system on the MMLoSo 2025 leaderboard (evaluation metric: $0.6 \times$BLEU$+0.4 \times$ chrF).

Ablation Study. Table 3 quantifies each component's contribution.

Qualitative Analysis. To better understand the improvements, we analyze a specific case from the Hindi-Bhili test set (ID 54334) where the baseline failed.

Case Study: Overcoming SMT Hallucinations

Input (Hindi): unhone kaha ki 2014 ke baad...
(Gloss: He said that after 2014...)

Baseline (SMT): **ki ki ki** 2014. baad...
× *Error: Severe stuttering and repetition at start.*

Hybrid System: **tinaye kedu ki** 2014 ne baad...
✓ *Correction: Fluent generation of "He said that".*

Analysis. Key insights: (1) **Complementary error modes**—SMT provides literal translations but with grammatical errors; NMT produces fluent output but hallucinates (public 302.08 vs private 166.47 confirms overfitting). (2) **MBR mitigates errors**—consensus selection adds +8.06 points over NMT-only. (3) **RAG excels in redundant domains**—contributes +11.84 points; exact matches handle 8% of test data with perfect accuracy. (4) **Post-processing is critical**—script-aware digit normalization adds +2.45 points for Indic languages.

5 Conclusion

We presented a hybrid translation system for the MMLoSo 2025 Shared Task, achieving 2nd place on the leaderboard with a score of 186.37. Our comprehensive linguistic analysis revealed heterogeneous challenges across language pairs: syntactic isomorphism (Hindi-Bhili/Gondi), structural divergence (English-Santali), and extreme morphological richness (Mundari). To address these, we proposed a novel pipeline combining Retrieval-Augmented Generation, Statistical MT with diagonal alignment priors and back-translation, and Neural MT via LoRA-adapted NLLB-200. Minimum Bayes-Risk decoding effectively synthesizes consensus translations from diverse hypotheses, mitigating complementary error modes.

Our ablation studies demonstrate that each component contributes substantially: MBR improves over NMT-only by +8 points, RAG adds +12 points, and post-processing contributes +2.5 points. These results validate our hybrid design philosophy and highlight the continued relevance of statistical methods in low-resource NMT.

Future Work. Promising directions include: (1) exploring iterative pseudo-labeling with

Method	Public Score	Private Score
Baselines		
Dice Coefficient (Lexical)	158.84	140.32
IBM Model 1 (SMT)	182.53	148.68
Intermediate Systems		
SMT + Back-Translation + MBR	193.26	153.91
NLLB-LoRA (Neural Only)	302.08	166.47
NLLB-LoRA + SMT + MBR	306.56	174.53
Final Hybrid System	**311.61**	**186.37**

Table 2: Comparison of system performance. The Final Hybrid System includes RAG, Ensemble, and Post-processing.

System Configuration	Score
NLLB-LoRA only	166.47
+ SMT ensemble	170.21
+ MBR reranking	174.53
+ RAG (Exact Match)	180.14
+ RAG (Fuzzy Match)	183.92
+ Post-processing (Digit mapping)	**186.37**

Table 3: Ablation study showing incremental contributions.

confidence-based filtering, (2) integrating subword-level MBR to better handle morphological variation, (3) developing language-pair-specific adapters to address structural heterogeneity, and (4) investigating cross-lingual transfer from related high-resource languages (e.g., Marathi for Gondi).

Limitations

While our system achieves competitive performance, several limitations warrant discussion:

Domain Specificity. Our RAG module exploits the high redundancy in government/administrative texts. Performance may degrade on out-of-domain data (e.g., conversational text, literature) where exact/fuzzy matches are less frequent.

Computational Cost. The hybrid pipeline requires running both SMT and NMT inference, increasing latency by approximately 2.5× compared to NMT-only. This may limit deployment in resource-constrained scenarios.

Error Propagation. The MBR reranking relies on BLEU and chrF as utility functions. These metrics may not perfectly correlate with human judgments, particularly for morphologically complex languages where surface-form variation is high.

Language Coverage. Our analysis focuses on four specific tribal languages. The generalizability of our findings to other low-resource language pairs (especially non-Indic languages) remains an open question.

Ethical Considerations. Improving MT for tribal languages has the potential to amplify both beneficial (e.g., access to government services) and harmful (e.g., loss of linguistic diversity) societal impacts. Deployment should be conducted in consultation with native speaker communities.

Acknowledgments

Include acknowledgments for funding sources, colleagues who provided feedback, or any other support you received.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (NAACL-HLT)*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation (NMT)*, pages 56–60, Vancouver, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 169–176, Boston, Massachusetts. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

MMLoSo Organizers. 2025. MMLoSo 2025 shared task: Multimodal models for low-resource contexts and social impact. `https://www.kaggle.com/competitions/mmloso2025`. To appear.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

A Detailed System Architecture

Our final system architecture is a multi-stage pipeline designed to maximize robustness and accuracy. The complete workflow is described below:

1. **Preprocessing**: All input sentences undergo normalization (NFKC) and whitespace standardization.

2. **Retrieval-Augmented Generation (RAG)**:
   - **Exact Match**: We check if the source sentence exists verbatim in the training data. If found, the corresponding target is returned immediately.
   - **Fuzzy Match**: We search for training sentences with a normalized edit distance of ≤ 1 character. This handles minor variations in punctuation or spacing.

3. **Hybrid Generation (if RAG fails)**:
   - **SMT Branch**: The input is processed by our IBM Model 1 system (enhanced with diagonal prior and back-translation). We generate the top-5 hypotheses using beam search.
   - **NMT Branch**: The input is processed by the NLLB-200-Distilled-600M model (fine-tuned with LoRA). We generate the top-10 hypotheses using beam search with a temperature of 1.0.

4. **Minimum Bayes-Risk (MBR) Reranking**:
   - We pool the hypotheses from both branches ($N = 15$).
   - We compute the utility score for each hypothesis against all others using the metric: $U(h) = 0.6 \times \text{BLEU}(h, h') + 0.4 \times \text{chrF}(h, h')$.
   - The hypothesis with the highest average utility is selected.

5. **Post-Processing**:
   - **Digit Mapping**: For Indic target languages (Hindi, Bhili, Gondi, Mundari), we map Latin digits (0-9) to Devanagari digits.
   - **Entity Preservation**: We verify that all URLs and email addresses present in the source are preserved in the target. If missing, they are appended.

B Hyperparameters and Configuration

We provide the detailed hyperparameters used for our best-performing models.

Parameter	Value
NLLB-200 (LoRA)	
Base Model	`nllb-200-distilled-600M`
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	`[q_proj, v_proj, k_proj, out_proj, fc1, fc2]`
Learning Rate	2×10^{-4}
Batch Size	16
Epochs	3
Quantization	8-bit (Int8)
SMT (IBM Model 1)	
EM Iterations	6
Diagonal Prior (λ_{diag})	4.0
Smoothing	Kneser-Ney (3-gram)
Back-Translation Rounds	3
MBR Decoding	
Candidate Pool Size	15 (5 SMT + 10 NMT)
Utility Function	$0.6 \cdot \text{BLEU} + 0.4 \cdot \text{chrF}$

Table 4: Hyperparameters for NMT, SMT, and MBR components.

C Detailed Experiment History

Table 5 lists the complete history of our experiments, showing the evolution from simple baselines to the final hybrid system.

D Linguistic Analysis Details

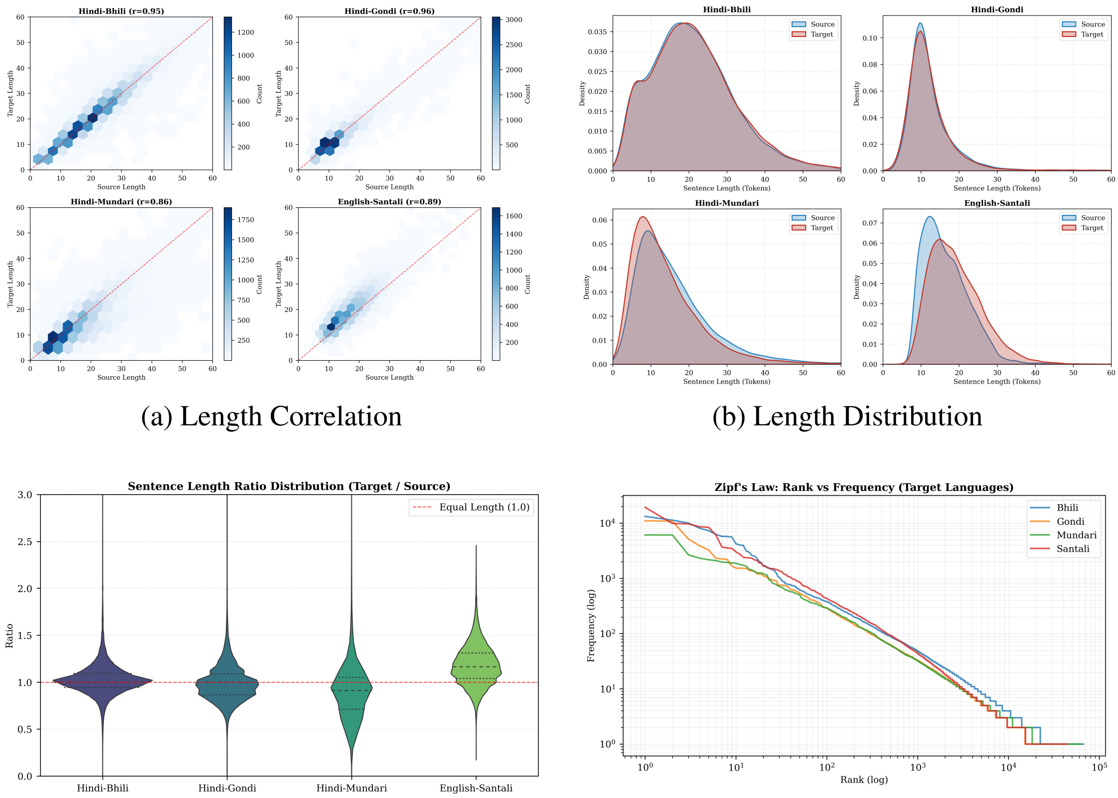
We performed a detailed analysis of the dataset characteristics to inform our model choices. Key observations from our analysis:

- **Isomorphism**: Hindi-Bhili and Hindi-Gondi are highly isomorphic (length correlation $r \geq 0.95$), with nearly identical sentence length ratios (≈ 1.00), justifying the use of SMT for these pairs.

- **Morphological Richness**: Mundari exhibits the highest Type-Token Ratio (TTR = 0.22), more than double that of Hindi, indicating extreme morphological complexity and data sparsity. This necessitated the use of Back-Translation for vocabulary expansion.

- **Structural Divergence**: English-Santali shows the lowest length correlation ($r = 0.89$) and a high length ratio (≈ 1.18), reflecting Santali's agglutinative morphology, suggesting that NMT is more suitable than SMT for this pair.

Visualizations of these characteristics are provided in Figure 2.



(a) Length Correlation

(b) Length Distribution

(c) Length Ratios

(d) Zipf's Law Analysis

Figure 2: Exploratory Data Analysis. (a) Hexbin plots showing strong isomorphism for Hindi-Bhili/Gondi. (b) KDE plots showing distribution overlap. (c) Violin plots of target/source length ratios. (d) Zipf's law plots confirming natural language properties.

ID	Method Description	Public	Private
	Phase 1: Statistical Baselines		
ML0	Dice Coefficient (Word-by-word, No LM)	158.84	140.32
ML5	IBM Model 1 + Word LM	182.53	148.68
ML1	IBM1 (Diag Prior) + KN LM + Char LM	175.83	143.91
Exp 3	**IBM1 (Diag) + Back-Translation + MBR**	**193.26**	**153.91**
	Phase 2: Neural Methods (NLLB)		
LLM0	NLLB LoRA + Dice Fallback (Early Hybrid)	171.64	161.10
LLM2	NLLB LoRA (Standard Fine-tuning)	302.08	166.47
LLM5	**NLLB LoRA + SMT + MBR (Best Single NMT)**	**306.56**	**174.53**
	Phase 3: Final Hybrid System		
Final	**RAG + NLLB-LoRA + SMT + MBR + Post-Proc**	**311.61**	**186.37**

Table 5: Complete experiment history showing the progression of Public and Private leaderboard scores.