

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) ($\text{lr} = 2e-4$), 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).

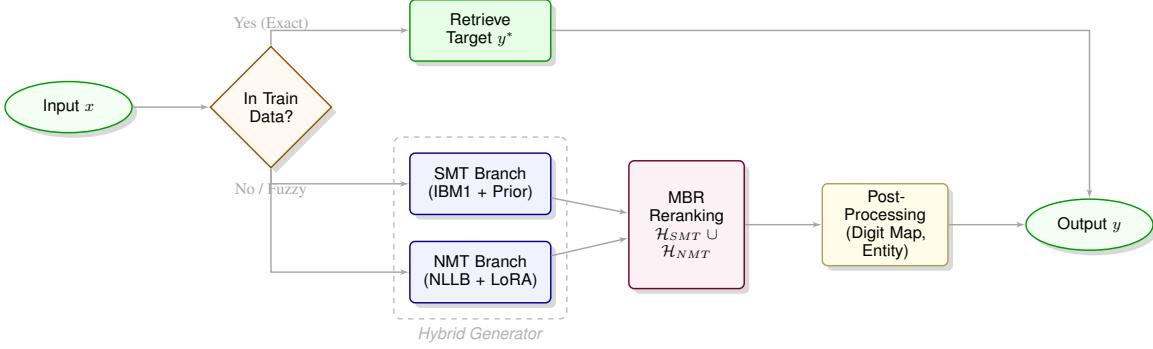


Figure 1: Architecture of our Hybrid Retrieval-Augmented Ensemble. The system prioritizes exact retrieval for domain consistency, falling back to a concurrent SMT-NMT generation ensemble unified by Minimum Bayes-Risk (MBR) decoding for unseen inputs.

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 $0.6 \times \text{BLEU}$ (Papineni et al., 2002) $+ 0.4 \times \text{chrF}$ (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: $0.6 \times \text{BLEU} + 0.4 \times \text{chrF}$).

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

160
161
162
163
164
165
166
167
168
169
170

171

172
173
174
175

176
177

178
179
180
181

182

183
184

185
186
187
188
189
190
191
192
193
194

195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210

211
212
213
214
215
216
217

Method	Public Score	Private Score
<i>Baselines</i>		
Dice Coefficient (Lexical)	158.84	140.32
IBM Model 1 (SMT)	182.53	148.68
<i>Intermediate Systems</i>		
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 $2.5\times$ 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*

283	Maja Popović.	2015. <i>chrF: character n-gram F-score for automatic MT evaluation</i> . In <i>Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)</i> , pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.	338
284	Rico Sennrich, Barry Haddow, and Alexandra Birch.		339
285	2016. <i>Improving neural machine translation models with monolingual data</i> . In <i>Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)</i> , pages 86–96, Berlin, Germany. Association for Computational Linguistics.	340	
286			341
287			342
288			343
289			344
290			345
291			346
292			347
293			348
294			349
295	A Detailed System Architecture		349
296	Our final system architecture is a multi-stage		350
297	pipeline designed to maximize robustness and accuracy. The complete workflow is described below:		351
298			352
299	1. Preprocessing: All input sentences undergo		353
300	normalization (NFKC) and whitespace standardization.		354
301			355
302	2. Retrieval-Augmented Generation (RAG):		356
303			357
304	• Exact Match: We check if the source		358
305	sentence exists verbatim in the training		359
306	data. If found, the corresponding target		360
307	is returned immediately.		361
308	• Fuzzy Match: We search for training		362
309	sentences with a normalized edit distance		363
310	of ≤ 1 character. This handles minor		364
311	variations in punctuation or spacing.		365
312	3. Hybrid Generation (if RAG fails):		365
313			366
314	• SMT Branch: The input is processed		367
315	by our IBM Model 1 system (enhanced		368
316	with diagonal prior and back-translation).		369
317	We generate the top-5 hypotheses using		370
318	beam search.		371
319	• NMT Branch: The input is processed		372
320	by the NLLB-200-Distilled-600M model		373
321	(fine-tuned with LoRA). We generate the		374
322	top-10 hypotheses using beam search		375
323	with a temperature of 1.0.		376
324	4. Minimum Bayes-Risk (MBR) Reranking:		376
325			377
326	• We pool the hypotheses from both		378
327	branches ($N = 15$).		379
328	• We compute the utility score for each		380
329	hypothesis against all others using the		381
330	metric: $U(h) = 0.6 \times \text{BLEU}(h, h') + 0.4 \times \text{chrF}(h, h')$.		382
331	• The hypothesis with the highest average		383
332	utility is selected.		384
333			384
334			384
335			384
336			384
337			384

385 5. Post-Processing:

- 386 • **Digit Mapping:** For Indic target lan-
387 guages (Hindi, Bhili, Gondi, Mundari),
388 we map Latin digits (0-9) to Devanagari
389 digits.
- 390 • **Entity Preservation:** We verify that all
391 URLs and email addresses present in the
392 source are preserved in the target. If
393 missing, they are appended.

394 B Hyperparameters and Configuration

395 We provide the detailed hyperparameters used for
396 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.

397 C Detailed Experiment History

398 Table 5 lists the complete history of our experi-
399 ments, showing the evolution from simple base-
400 lines to the final hybrid system.

401 D Linguistic Analysis Details

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

- 405 • **Isomorphism:** Hindi-Bhili and Hindi-Gondi
406 are highly isomorphic (length correlation $r \geq$
407 0.95), with nearly identical sentence length
408 ratios (≈ 1.00), justifying the use of SMT for
409 these pairs.
- 410 • **Morphological Richness:** Mundari exhibits
411 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.

- 412 • **Structural Divergence:** English-Santali
413 shows the lowest length correlation ($r = 0.89$)
414 and a high length ratio (≈ 1.18), reflecting
415 Santali’s agglutinative morphology, suggest-
416 ing that NMT is more suitable than SMT for
417 this pair.

Visualizations of these characteristics are pro-
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
8010
8011
8012
8013
8014
8015
8016
8017
8018
8019
8020
8021
8022
8023
8024
8025
8026
8027
8028
8029
8030
8031
8032
8033
8034
8035
8036
8037
8038
8039
8040
8041
8042
8043
8044
8045
8046
8047
8048
8049
8050
8051
8052
8053
8054
8055
8056
8057
8058
8059
8060
8061
8062
8063
8064
8065
8066
8067
8068
8069
8070
8071
8072
8073
8074
8075
8076
8077
8078
8079
8080
8081
8082
8083
8084
8085
8086
8087
8088
8089
8090
8091
8092
8093
8094
8095
8096
8097
8098
8099
80100
80101
80102
80103
80104
80105
80106
80107
80108
80109
80110
80111
80112
80113
80114
80115
80116
80117
80118
80119
80120
80121
80122
80123
80124
80125
80126
80127
80128
80129
80130
80131
80132
80133
80134
80135
80136
80137
80138
80139
80140
80141
80142
80143
80144
80145
80146
80147
80148
80149
80150
80151
80152
80153
80154
80155
80156
80157
80158
80159
80160
80161
80162
80163
80164
80165
80166
80167
80168
80169
80170
80171
80172
80173
80174
80175
80176
80177
80178
80179
80180
80181
80182
80183
80184
80185
80186
80187
80188
80189
80190
80191
80192
80193
80194
80195
80196
80197
80198
80199
80200
80201
80202
80203
80204
80205
80206
80207
80208
80209
80210
80211
80212
80213
80214
80215
80216
80217
80218
80219
80220
80221
80222
80223
80224
80225
80226
80227
80228
80229
80230
80231
80232
80233
80234
80235
80236
80237
80238
80239
80240
80241
80242
80243
80244
80245
80246
80247
80248
80249
80250
80251
80252
80253
80254
80255
80256
80257
80258
80259
80260
80261
80262
80263
80264
80265
80266
80267
80268
80269
80270
80271
80272
80273
80274
80275
80276
80277
80278
80279
80280
80281
80282
80283
80284
80285
80286
80287
80288
80289
80290
80291
80292
80293
80294
80295
80296
80297
80298
80299
802100
802101
802102
802103
802104
802105
802106
802107
802108
802109
802110
802111
802112
802113
802114
802115
802116
802117
802118
802119
802120
802121
802122
802123
802124
802125
802126
802127
802128
802129
802130
802131
802132
802133
802134
802135
802136
802137
802138
802139
802140
802141
802142
802143
802144
802145
802146
802147
802148
802149
802150
802151
802152
802153
802154
802155
802156
802157
802158
802159
802160
802161
802162
802163
802164
802165
802166
802167
802168
802169
802170
802171
802172
802173
802174
802175
802176
802177
802178
802179
802180
802181
802182
802183
802184
802185
802186
802187
802188
802189
802190
802191
802192
802193
802194
802195
802196
802197
802198
802199
802200
802201
802202
802203
802204
802205
802206
802207
802208
802209
802210
802211
802212
802213
802214
802215
802216
802217
802218
802219
802220
802221
802222
802223
802224
802225
802226
802227
802228
802229
802230
802231
802232
802233
802234
802235
802236
802237
802238
802239
802240
802241
802242
802243
802244
802245
802246
802247
802248
802249
802250
802251
802252
802253
802254
802255
802256
802257
802258
802259
802260
802261
802262
802263
802264
802265
802266
802267
802268
802269
802270
802271
802272
802273
802274
802275
802276
802277
802278
802279
802280
802281
802282
802283
802284
802285
802286
802287
802288
802289
802290
802291
802292
802293
802294
802295
802296
802297
802298
802299
802300
802301
802302
802303
802304
802305
802306
802307
802308
802309
802310
802311
802312
802313
802314
802315
802316
802317
802318
802319
802320
802321
802322
802323
802324
802325
802326
802327
802328
802329
802330
802331
802332
802333
802334
802335
802336
802337
802338
802339
802340
802341
802342
802343
802344
802345
802346
802347
802348
802349
802350
802351
802352
802353
802354
802355
802356
802357
802358
802359
802360
802361
802362
802363
802364
802365
802366
802367
802368
802369
802370
802371
802372
802373
802374
802375
802376
802377
802378
802379
802380
802381
802382
802383
802384
802385
802386
802387
802388
802389
802390
802391
802392
802393
802394
802395
802396
802397
802398
802399
802400
802401
802402
802403
802404
802405
802406
802407
802408
802409
802410
802411
802412
802413
802414
802415
802416
802417
802418
802419
802420
802421
802422
802423
802424
802425
802426
802427
802428
802429
802430
802431
802432
802433
802434
802435
802436
802437
802438
802439
802440
802441
802442
802443
802444
802445
802446
802447
802448
802449
802450
802451
802452
802453
802454
802455
802456
802457
802458
802459
802460
802461
802462
802463
802464
802465
802466
802467
802468
802469
802470
802471
802472
802473
802474
802475
802476
802477
802478
802479
802480
802481
802482
802483
802484
802485
802486
802487
802488
802489
802490
802491
802492
802493
802494
802495
802496
802497
802498
802499
802500
802501
802502
802503
802504
802505
802506
802507
802508
802509
802510
802511
802512
802513
802514
802515
802516
802517
802518
802519
802520
802521
802522
802523
802524
802525
802526
802527
802528
802529
802530
802531
802532
802533
802534
802535
802536
802537
802538
802539
802540
802541
802542
802543
802544
802545
802546
802547
802548
802549
802550
802551
802552
802553
802554
802555
802556
802557
802558
802559
802560
802561
802562
802563
802564
802565
802566
802567
802568
802569
802570
802571
802572
802573
802574
802575
802576
802577
802578
802579
802580
802581
802582
802583
802584
802585
802586
802587
802588
802589
802590
802591
802592
802593
802594
802595
802596
802597
802598
802599
802600
802601
802602
802603
802604
802605
802606
802607
802608
802609
802610
802611
802612
802613
802614
802615
802616
802617
802618
802619
802620
802621
802622
802623
802624
802625
802626
802627
802628
802629
802630
802631
802632
802633
802634
802635
802636
802637
802638
802639
802640
802641
802642
802643
802644
802645
802646
802647
802648
802649
802650
802651
802652
802653
802654
802655
802656
802657
802658
802659
802660
802661
802662
802663
802664
802665
802666
802667
802668
802669
802670
802671
802672
802673
802674
802675
802676
802677
802678
802679
802680
802681
802682
802683
802684
802685
802686
802687
802688
802689
802690
802691
802692
802693
802694
802695
802696
802697
802698
802699
802700
802701
802702
802703
802704
802705
802706
802707
802708
802709
802710
802711
802712
802713
802714
802715
802716
802717
802718
802719
802720
802721
802722
802723
802724
802725
802726
802727
802728
802729
802730
802731
802732
802733
802734
802735
802736
802737
802738
802739
802740
802741
802742
802743
802744
802745
802746
802747
802748
802749
802750
802751
802752
802753
802754
802755
802756
802757
802758
802759
802760
802761
802762
802763
802764
802765
802766
802767
802768
802769
802770
802771
802772
802773
802774
802775
802776
802777
802778
802779
802780
802781
802782
802783
802784
802785
802786
802787
802788
80

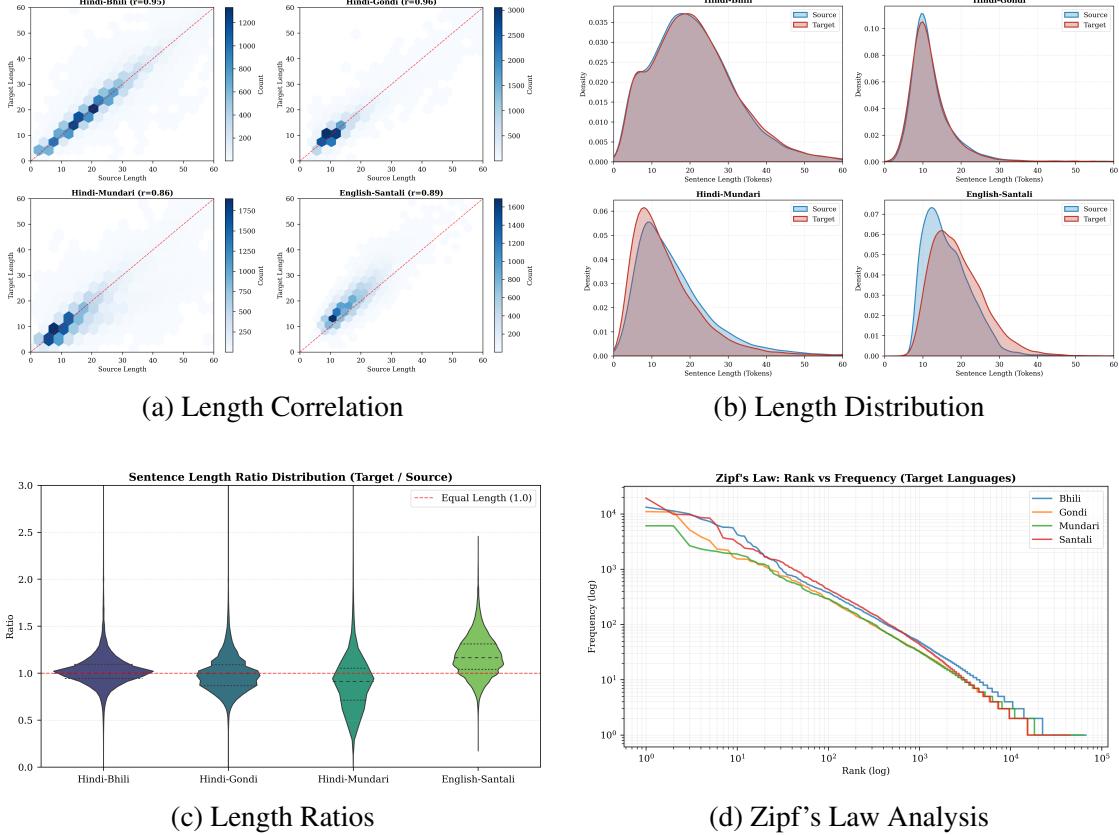


Figure 2: Exploratory Data Analysis. (a) Hexbin plots showing strong isomorphism for Hindi-Bhili/Gondi. (b) KDE plots showing distribution overlap. (c) Violin plots of target/source length ratios. (d) Zipf’s law plots confirming natural language properties.

ID	Method Description	Public	Private
<i>Phase 1: Statistical Baselines</i>			
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
<i>Phase 2: Neural Methods (NLLB)</i>			
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
<i>Phase 3: Final Hybrid System</i>			
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.