

From Folklore to Fine-Tuning: An Authentic Angika-Hindi Parallel Corpus and Warm-Initialized Translation Baseline

Jyoti Kumari

Department of Linguistics,
Banaras Hindu University, U.P, India
jyoti@bhu.ac.in

Abstract

While Massively Multilingual Machine Translation (MMT) models and Large Language Models (LLMs) have achieved remarkable success for high-resource languages, extremely low-resource languages (ELRLs) remain severely marginalized. Angika, a Bihari language spoken by millions, lacks digitized parallel corpora and is absent from major LLM training distributions and MMT benchmarks like NLLB-200. This paper presents a manually constructed, high-quality Angika-Hindi parallel corpus (10,068 pairs). This corpus is sourced from digitized physical folklore, preserving authentic *Deshaj* (indigenous) vocabulary and idioms. To build an effective translation system, the NLLB-200 tokenizer was expanded and its embeddings were phylogenetically warm-initialized using Maithili, its closest supported relative. The resulting lightweight 600M-parameter model was comprehensively evaluated against eight state-of-the-art LLMs, including Gemini 2.5 Flash, GPT-4o and Llama 3.3 (70B). The findings reveal a generation-comprehension gap in modern LLMs: while models like Gemini 2.5 Flash achieve comparable comprehension scores (50.88 BLEU for Angika→Hindi), they suffer catastrophic lexical bleed during generation, peaking at only 13.95 BLEU for Hindi→Angika. In contrast, warm-initialized model achieves 30.68 BLEU for Hindi→Angika, demonstrating that adding a dedicated language token, whether initialized randomly or from a related language, is effective for unseen ELRLs compared to naive prompting.

1 Introduction

The rapid advancement of deep learning has ushered in a new era of natural language processing (NLP), largely driven by Massively Multilingual Machine Translation (MMT) models (Costa-Jussà et al., 2022; Xue et al., 2020) and instruction-tuned Large Language Models (LLMs) (Üstün

et al., 2024; Grattafiori et al., 2024). However, this progress is uneven. The vast majority of the world’s languages are categorized as extremely low-resource languages (ELRLs) (Koehn and Knowles, 2017). Because modern architectures rely on large-scale, web-crawled digital footprints (Kreutzer et al., 2022), languages without a digital presence are frequently excluded from the benefits of modern NLP.

A prime example of this digital divide is Angika, an Eastern Indo-Aryan language belonging to the Bihari subgroup of the Indo-Aryan language family (Masica, 1991). It is spoken predominantly in the historical Anga region of eastern Bihar, particularly in the districts of Bhagalpur, Banka, Katihar, Purnia, Madhepura, Saharsa, Munger and Sheikhpura. The language is also spoken in parts of Jharkhand, West Bengal and neighbouring areas of Nepal (Regmi, 2017). According to the *Mother Tongue Survey of India* (2023), Angika continues to serve as an important medium of everyday communication across these regions. The language is primarily written in the Devanagari script, although historical sources indicate the use of Anga Lipi and Kaithi for writing Angika. George A. Grierson(1903) classified Angika within the Bihari languages in the *Linguistic Survey of India*, where it was referred to as “Chika Chiki.” Since the 1961 Census of India, the language has been recorded under the name Angika. Despite its well-established linguistic identity and rich heritage of folklore and literature, Angika remains largely underrepresented in the digital NLP ecosystem, with limited standardized digital corpora, syntactic resources and other computational linguistic resources.

Recent research has begun to explore translation for unseen ELRLs, making vital strides in expanding language coverage. For instance, recent work on Angika translation, such as the SrcMix approach (Kumar et al., 2026), introduced a valu-

able dataset of 6,192 pairs by translating standard English NLLB seed corpora into Angika. This methodology is effective for establishing foundational MT baselines and ensuring broad compatibility with existing multilingual benchmarks.

Approaching the low-resource challenge from a complementary direction, this work seeks to digitize the language exactly as it is natively authored. While translating standardized English text into Angika provides an excellent structural baseline, natively written texts naturally contain deep cultural nuances, vivid metaphors and *Deshaj* (indigenous, rural) vocabulary that rarely appear in global seed corpora. To capture the linguistic richness, a rigorous print-to-digital pipeline was designed. By sourcing physical Angika storybooks and folklore, performing Optical Character Recognition (OCR) and meticulously correcting typographic artifacts, an authentic, manually translated Angika–Hindi parallel corpus comprising 10,068 sentence pairs was constructed.

Furthermore, integrating an unseen language into an existing translation pipeline poses significant architectural challenges. Modern LLMs operate on fixed subword tokenizers that lack representation for unseen ELRLs, causing them to hallucinate or default to related high-resource languages, a phenomenon known as *lexical bleed* (Bapna et al., 2022). To address this, the nllb-200-distilled-600M architecture is adapted. A dedicated language token for Angika is explicitly registered (anp_Deva) within its tokenizer and explore a *phylogenetic warm-initialization* strategy, extracting the latent representations of Maithili (mai_Deva) as a structural anchor.

To rigorously evaluate the proposed approach, the model was benchmarked against eight state-of-the-art LLMs, ranging from 12B to 235B parameters, on the held-out test set. The results reveal a stark asymmetry in modern LLM capabilities. While massive models demonstrate strong cross-lingual comprehension (Angika→Hindi), they fail completely at generative translation (Hindi→Angika). In contrast, the fine-tuned 600M model outperforms models 100x its size by over +16 BLEU points in the generative direction. While the phylogenetic warm-initialization offers a statistically narrow performance edge over random initialization, the experiment demonstrates that the registration of a dedicated language token, coupled with focused fine-tuning, presents a stable baseline for unseen ELRLs, far exceeding the

generative capacity of massive zero-shot LLMs.

In summary, contributions are as follows:

1. **Authentic Data Creation:** This paper presents the first high-quality, manually translated parallel corpus for Angika–Hindi(10,068 pairs), digitized directly from physical regional folklore to preserve indigenous linguistic structures.
2. **Token Registration and Initialization:** The NLLB-200 architecture is expanded for an unseen ELRL by registering a dedicated language token. The experiments demonstrate that although phylogenetic warm-initialization provides a useful heuristic, the model remains robust across different initialization strategies; the critical intervention preventing lexical bleed is the dedicated token itself.
3. **Comprehensive LLM Benchmarking:** This paper provides the first systematic evaluation of state-of-the-art LLMs(Gemini 2.5, GPT-4o, Claude Sonnet 4, Llama 3.3) on Angika, revealing a critical generation-comprehension gap that the targeted fine-tuning methodology overcomes.

The complete Angika-Hindi parallel corpus, codes and predictions are publicly available on GitHub repository at https://github.com/jyoti0502/angika_anp-hin and on HuggingFace at <https://huggingface.co/jyoti0/nllb-anp-ft/tree/main>.

2 Corpus Creation and Digitization

The primary bottleneck for Angika NLP is the absolute scarcity of digital text. To construct Angika-Hindi parallel corpus, a rigorous print-to-digital pipeline was designed moving from physical regional literature to a clean, digital parallel corpus.

2.1 Sourcing and Digital Acquisition

Because Angika has a limited digital presence and lacks sufficient web-crawled resources, data collection relied primarily on printed materials. Regional Angika literature was physically collected from the Department of Angika at Tilka Manjhi Bhagalpur University, Bihar, India. The curated texts represent prominent regional literature, including *Angika Anu Kahani*, *Gulabo*, *Angika*

Kahni Sangrah, Kharsup, Sukari and Chandan Van Me Rakas.

To convert the collected printed materials into a digitized corpus, the pages were scanned and processed using Google Lens¹. The OCR system first captured the visual content and identified regions containing text by detecting and segmenting words, lines and paragraphs. Although modern OCR systems generally perform well on standardized Devanagari texts such as Hindi (Smith, 2007), Google Lens encountered considerable difficulties with regional Angika typography. Consequently, the initial OCR output contained frequent errors, averaging approximately 1–2 corrupted words per line of text. The extracted text was therefore copied into a Microsoft Word document and manually reviewed and corrected to ensure accuracy. The corrected Angika text was subsequently transferred to an Excel spreadsheet, where corresponding Hindi translations were manually added, resulting in the creation of the Angika–Hindi parallel corpus.

2.2 OCR Post-Processing and Noise Reduction

To transform the raw OCR output into a linguistically sound dataset, a manual multi-stage review and sanitization pipeline was enacted:

1. **Typographic Correction:** The corpus was manually corrected for thousands of systematic errors, including redundant nuqtas (दौड़ी → दौड़ी), extraneous diacritics (थोड़ों → थोड़ी) and character misrecognitions (e.g., रड़ → रड़, भूमै → झूमै).
2. **Metadata and Artifact Removal:** Printed book elements that disrupt natural sentence flow were stripped. This included the removal of parenthetical page numbers (e.g., - (21)), indexing markers (e.g., -52-148), decorative section breaks (e.g., ○ ○ ○) and chapter headers (e.g., अनिरुद्ध प्रसाद विमल / १२, शुकरी / १).
3. **Non-Narrative Filtering:** Intralingual elements that do not map to parallel translation models were discarded. Embedded folk songs (e.g., गावै लागलो छे लै-किनयाय-मिनमाय झींगा झोर, किनयाय माय के लै गेलै चोर।) and excessive emotive punctuation used as text fillers

(e.g., हा..... or लेकिन) were deleted to maintain strict sentence-level parallel alignment.

2.3 Translation and Verification Protocol

Following text sanitization, the resulting corpus was manually translated into standard Hindi. To maintain strict quality control under the constraints of a single-translator project, a two-party verify-and-check protocol was utilized. The primary translation of the 10,068 pairs was executed by the primary researcher. Subsequently, the compiled parallel text was delivered to a native Angika and Hindi speaker, who reviewed the sentence pairs. This secondary verification resolved regional variations, verified the preservation of indigenous *Deshaj* idioms and ensured the syntactic integrity of the Hindi mappings.

2.4 Manual Translation Challenges

During translation, several profound linguistic divergences between the two languages required standardized guidelines to ensure corpus consistency:

- **Gender Neutrality in Verbs:** Standard Hindi verbs heavily inflect based on gender (e.g., *vah jaata hai* vs. *vah jaati hai*). In contrast, Angika is largely gender-neutral in its verbal morphology. The Angika phrase *u jaay chhai* is used identically for both “he goes” and “she goes”. Contextual cues were required to map gender correctly into Hindi.
- **Deshaj (Indigenous) Vocabulary:** Angika possesses a rich vocabulary of rural and agricultural words lacking direct equivalents in standard Hindi. For example, *phulaais* refers to a specific religious ritual performed during a puja. Translating such *Deshaj* terms required strict descriptive guidelines, utilizing multi-word descriptive expansions in Hindi rather than default phonetic transliteration.
- **Idiomatic Expressions:** Angika folklore relies on vivid metaphors that do not map directly to Hindi. For instance, the temporal idiom *garmi mein te baans bhari surujo ke rahatan u aabi jaay chhai* (literally referencing the sun’s position relative to bamboo) means “while the sun is still out.” This was translated to the standard Hindi *garmi mein to din rahte hi vah aa jaata hai*.

¹<https://lens.google.com/>

2.5 Dataset Statistics

Following the manual translation and strict secondary verification, the final deduplicated dataset consists of 10,068 high-quality parallel sentence pairs. To ensure robust evaluation and prevent data leakage, the corpus was partitioned into a Training Set (7,651 pairs), a Development Set (402 pairs; utilized for per-epoch validation) and an untouched Test Set (2,015 pairs).

3 Methodology

The facebook/nllb-200-distilled-600M architecture (Costa-Jussà et al., 2022), an encoder-decoder Transformer optimized for low-resource translation, is adapted and evaluated against both ablated baselines and modern LLMs.

3.1 Dataset Partitioning and Deduplication

The 10,068 Angika-Hindi parallel corpus were randomly shuffled and divided into training and test sets using an 80:20 split. This resulted in 8,053 sentence pairs for training and 2,015 pairs for testing. During model fine-tuning, the training set was further divided into a training subset of 7,651 pairs and a development set of 402 pairs, which was used exclusively for validation at the end of each training epoch. The test set remained completely untouched throughout the training and model selection process.

Because the corpus was compiled from multiple books and stories, a random sentence-level split may introduce potential document-level overlap between the training and test sets. For example, adjacent passages, recurring expressions, character names, narrative patterns and author-specific writing styles may occur in both partitions and could potentially lead to inflated evaluation scores. To reduce this risk, the corpus was examined for exact duplicate sentence pairs. Such duplicates were removed where identified. However, the present dataset follows a sentence-level random split rather than a document-disjoint split by book or story.

3.2 Tokenizer Registration

NLLB utilizes language-specific special tokens to guide generation. It is important to clarify that no new SentencePiece vocabulary was trained specifically for Angika. Because Angika shares the Devanagari script and significant morphological overlap with existing supported languages (e.g.,

Hindi and Maithili), the approach relies entirely on NLLB’s existing subword vocabulary.

The tokenizer was expanded by registering a single target-language ID token `anp_Deva`, via the `additional_special_tokens` configuration. Consequently, the embedding matrix of the base model was resized by a single row to accommodate the new token ID.

3.3 Phylogenetic Warm-Initialization

Randomly initializing a new language-ID embedding vector for a low-resource language often causes poor latent alignment (Chronopoulou et al., 2020). To circumvent this, a **phylogenetic warm-initialization** strategy was designed.

Angika and Maithili both belong to the Bi-hari sub-group of the Indo-Aryan language family, sharing significant syntactic structure. Because NLLB-200 already features robust representations for Maithili (`mai_Deva`), extracted the weights of the `mai_Deva` token and cloned them directly into both the input and output embedding rows of the newly created `anp_Deva` token.

While this initializes only a single token embedding, copying it provides a critical structural prior. It primes the model’s decoder to map latent representations into the Indo-Aryan morphological space associated with Maithili, efficiently guiding the model to combine existing Devanagari subwords to construct Angika morphology rather than learning the language ID embedding from a blank slate.

3.4 Fine-Tuning Configuration

A bidirectional training objective was constructed by concatenating the Hindi→Angika and Angika→Hindi pairs. During experimentation, encountered a known auto-shift issue within the HuggingFace framework (Wolf et al., 2020) when training M2M architectures with label smoothing. To bypass this, implemented a custom `DataCollatorForSeq2Seq` that explicitly intercepts the batch and applies `shift_tokens_right` to build the `decoder_input_ids` manually.

To facilitate training of the 600M parameter model on a single NVIDIA T4 GPU (16 GB VRAM), the Adafactor optimizer (Shazeer and Stern, 2018) was utilized, which maintains a sub-linear memory footprint. The model was fine-tuned for 5 epochs with a learning rate of 3×10^{-5} , an effective batch size of 16 (batch size 4 with 4 accumulation steps), label smoothing of 0.1 and a 5%

warmup ratio. The best checkpoint was selected based on chrF++ evaluations on the development set.

3.5 LLM Benchmarking Setup

To contextualize the performance of the customized MMT approach, translation capabilities were evaluated across a diverse suite of state-of-the-art LLMs, ranging from 12B to 235B parameters. The benchmark includes both closed-source proprietary models accessed through APIs and open-weight architectures.

For proprietary models, **GPT-4o** (OpenAI, 2024), Anthropic’s **Claude Sonnet 4** (Anthropic, 2025) and Google’s **Gemini 2.5 Flash** (Comanici et al., 2025) was evaluated.

For open-weights foundation models, Meta’s **Llama 3.3 (70B)** (Mistral AI, 2026), Alibaba’s **Qwen3 (235B)** (Yang et al., 2025), **DeepSeek V3.2** (Liu et al., 2024), **Mistral Medium 3.5** (Mistral AI, 2026) and Google’s **Gemma 3 (12B)** (Team et al., 2024) was evaluated.

Zero-shot prompting was employed to simulate extreme low-resource conditions without in-context examples. The models were instructed via a standard system prompt: “*Translate the following [Source Language] sentence to the [Target Language] language in Devanagari script: [Sentence].*” The raw generated text was extracted for automated evaluation against the gold-standard test set. All automated metrics were computed using the SacreBLEU implementation (Post, 2018), specifically utilizing BLEU and chrF++ (Popović, 2015).

3.6 Ablation Study Design

To isolate and quantify the impact of the phylogenetic transfer strategy and to account for training variance, An ablation study was conducted with the training data and hyperparameters held constant across three independent random seeds. The proposed Maithili (mai_Deva) initialization was compared against two baselines:

1. **Random Initialization:** The anp_Deva embeddings are initialized from a standard normal distribution.
2. **Hindi Initialization:** The anp_Deva embeddings are cloned from the standard Hindi (hin_Deva) token, testing whether proximity to the dominant anchor language outper-

forms geographic and typological proximity (Maithili).

3.7 Human Evaluation Setup

While automated metrics provide consistent benchmarking, they can struggle to capture semantic nuances in morphologically rich ELRLs. To complement the automated evaluation, a targeted human evaluation was performed to assess the translation quality. A random sample of 100 test sentences was selected, with the source text presented alongside predictions from both the proposed NLLB model and the highest-performing LLM (Gemini 2.5 Flash). To eliminate system-order bias, the predictions were completely blinded and randomly shuffled for each sentence.

The evaluation was conducted by an independent linguist, a native Angika speaker who was deliberately excluded from the corpus creation and translation phases. The evaluator scored the blinded outputs on a 1–5 Likert scale for **Adequacy** (how much of the original meaning is preserved) and **Fluency** (how natural and grammatically correct the output is).

4 Results and Analysis

4.1 Comparison Against LLM Baselines

To evaluate the effectiveness of the targeted MMT approach, the fine-tuned 600M-parameter model was evaluated against a diverse set of state-of-the-art LLMs. The baselines range from 12B to 235B parameters, representing both closed-source APIs (Gemini, GPT-4o, Claude) and open-weights architectures (Llama, DeepSeek, Qwen).

As shown in Table 1, modern LLMs perform adequately in the Angika→Hindi direction, with Gemini 2.5 Flash achieving a BLEU score (50.88) comparable to the fine-tuned model. This indicates strong cross-lingual comprehension: these massive models can successfully map Angika vocabulary to their robust internal Hindi representations.

However, *all* LLMs suffer a catastrophic performance drop in the Hindi→Angika generative direction. Despite being orders of magnitude larger, the best-performing LLM achieves only 13.95 BLEU, compared to NLLB model’s 30.68 BLEU. This stark contrast highlights the fundamental limitation of massive LLMs for unseen ELRLs: lacking dedicated tokenizer representations and structural

Model	Hindi → Angika		Angika → Hindi	
	BLEU	chrF++	BLEU	chrF++
Proposed (NLLB 600M)	30.68	58.42	50.72	69.83
Gemini 2.5 Flash	13.95	44.44	50.88	70.77
Claude Sonnet 4	11.87	42.39	42.67	64.82
GPT-4o	11.18	40.37	45.22	66.62
DeepSeek V3.2	9.02	37.72	35.84	59.19
Llama 3.3 (70B)	8.37	38.14	30.54	56.17
Mistral Medium 3.5	6.27	33.30	32.36	57.01
Qwen3 (235B)	5.82	34.86	29.22	54.22
Gemma 3 (12B)	3.25	21.93	21.59	45.12

Table 1: Automated metric comparison on the 2,015-pair test set. 600M model vastly outperforms massive LLMs in the generative (Hindi→Angika) direction.

priors, they default to severe lexical bleed, out-putting standard Hindi rather than actual Angika.

4.2 Ablation Study: The Role of Initialization

To empirically evaluate phylogenetic transfer hypothesis, an ablation study was conducted holding the training data and hyperparameters constant. The Maithili (mai_Deva) initialization was compared against a random-weight initialization baseline and a standard Hindi (hin_Deva) baseline across both translation directions.

Initialization	Hindi → Angika		Angika → Hindi	
	BLEU	chrF++	BLEU	chrF++
Random Weights	30.50	58.39	50.31	69.46
Hindi (hin_Deva)	30.48	58.51	50.41	69.55
Maithili (mai_Deva)	30.68	58.42	50.72	69.83

Table 2: Ablation of token initialization strategies across both translation directions (5 epochs, 7.6K pairs).

As observed in Table 2, the proposed Maithili initialization yields the highest overall BLEU scores in both translation directions (30.68 and 50.72). However, a critical observation is that the performance margins between Maithili, Hindi and Random initializations are exceptionally narrow (≤ 0.2 BLEU and chrF++).

Because these fractional differences fall within the standard variance expected during LLM fine-tuning, these findings call for a more cautious interpretation of the initial hypothesis regarding phylogenetic transfer. While initializing from a typologically proximate language (Maithili) or an orthographically dominant language (Hindi) provides a safe, theoretically grounded starting point, the empirical data suggests that NLLB’s pre-existing multilingual representations are ro-

bust. The model learns the new anp_Deva token from scratch (Random initialization) nearly as effectively as it adapts a warm-initialized token.

Ultimately, this confirms that the critical factor preventing lexical bleed and enabling high-quality translation for unseen ELRLs is not necessarily *how* the token is initialized. Rather, the absolute performance gains are derived from the architectural registration of a dedicated language token coupled with focused, high-quality parallel fine-tuning.

4.3 Human Evaluation

To validate whether automated subword metrics align with true semantic quality in morphologically rich ELRLs (Freitag et al., 2022), A targeted human evaluation was conducted on a random sample of 100 test sentences, with the source text presented alongside blinded predictions from the fine-tuned NLLB model and the highest-performing LLM baseline(Gemini 2.5 Flash).

As outlined in Section 3.6, a native Angika speaker evaluated the blinded outputs on a 1–5 Likert scale for **Adequacy** (preservation of source meaning) and **Fluency** (grammatical naturalness).

System	Direction	Adequacy	Fluency	Overall
NLLB	Hin → Ang	3.96	3.87	3.92
Gemini 2.5 Flash	Hin → Ang	3.61	3.26	3.44
NLLB	Ang → Hin	3.94	3.95	3.95
Gemini 2.5 Flash	Ang → Hin	3.90	3.84	3.87

Table 3: Human evaluation results (1–5 Likert scale) averaged across 100 randomly sampled test predictions.

The human evaluation results (Table 3) validate the empirical superiority of the fine-tuned MMT approach over zero-shot frontier LLMs. Across both translation directions, the proposed NLLB

model achieves near-human quality, averaging approximately 3.95 out of 5.0.

In the critical generative direction (Hindi→Angika), NLLB comfortably outpaces Gemini 2.5 Flash (+0.35 Adequacy, +0.61 Fluency), with the difference in Fluency being pronounced. Interestingly, Gemini still achieves a relatively high Adequacy score (3.61) despite its low automated BLEU score (13.95). Qualitative inspection reveals this is a byproduct of *lexical bleed*: because bilingual evaluators naturally comprehend standard Hindi, they can decipher the LLM’s Hindi fallback vocabulary. However, the evaluator gave considerably lower scores for the LLM’s Fluency (3.26) due to persistent Maithili grammatical intrusions, incorrect gendered verb inflections and severe pronoun person-number misalignments. In contrast, NLLB model correctly follows Angika’s gender-neutral verb forms and its distinctive case markers.

4.4 Deep Linguistic Error Analysis

To understand the exact failure modes of both zero-shot LLMs and fine-tuned MMTs, The predictions were subjected to a detailed linguistic analysis, revealing four primary categories of divergence..

1. Lexical Bleed and Typological Interference:

When generating Angika, Gemini frequently hallucinates Maithili vocabulary and grammar. For instance, Gemini translates the Hindi *maa, tum kuch bolti kyon nahi ho?* using Maithili terms (*ahaan* and *kichhu*) and a Maithili verb structure (*bha chukal*), indicating that LLMs cluster low-resource Eastern Indo-Aryan languages together, causing severe cross-lingual pollution. Furthermore, both Gemini and NLLB suffer from standard Hindi retention, keeping words like *munaadi*, *phadfadaa* and *kal* unchanged rather than predicting the correct Angika equivalents. Conversely, in the Angika→Hindi direction, NLLB frequently transliterates *Deshaj* (indigenous) words rather than translating them (e.g., retaining the Angika *deh-gaat* instead of the Hindi *sharir-ang*).

2. Pronoun Misalignment and Nominal Inflection:

Angika utilizes specific agglutinative inflections that models struggle to map. In Hindi→Angika, NLLB fails to apply Angika’s unique emphasis and case markers to nouns. For example, Hindi *Sunita bhi* (Sunita also) should become *Sunito* in Angika; NLLB outputs *Sunita bhi*. Similarly, Hindi *Chandrakant ko* should inflect to

Chandrakanto ke, but NLLB outputs *Chandrakant ke*.

Gemini struggles profoundly with pronoun number and person mapping. It translates the singular Angika third-person pronoun *ven* (he/she) into the plural Hindi *ve* (they). Similarly, it translates the singular Angika first-person *hammen* (I) as the plural Hindi *hamne* (we), fundamentally altering the narrative subject. Furthermore, NLLB mistranslates the Hindi plural *hamaare* (our) to the Angika singular *hamro* (my), instead of the correct plural *hamra sini*.

3. Tense, Mood and Gender Resolution:

Angika verbs are largely gender-neutral, relying on context for resolution, whereas Hindi enforces verb-gender agreement. NLLB frequently forces incorrect Hindi gender rules onto Angika contexts; for a sentence clearly describing a male subject, NLLB outputs *kilittar banti* (a feminine Hindi verb conjugation).

Furthermore, models struggle to align grammatical mood and intent. Gemini translates the Angika future declarative *raksha karbai* (I will protect) into the Hindi imperative/suggestive *raksha karen* (please protect). Similarly, NLLB mistranslates the Hindi first-person future *karenge* (we will do) into the Angika third-person *kartai* (someone else will do).

4. Named Entity and Semantic Hallucinations:

Both models struggle with entity preservation. In Angika→Hindi, Gemini forcefully regularizes culturally specific names, translating the proper noun *Shukri* into the Hindi greeting *shukriya* (thank you) and reducing the name *Lalavatiya* to the generic Hindi noun *patni* (wife). NLLB similarly fails on hyphenated entities, hallucinating *alag-alag log* (different people) when presented with the names *Algu-Jumman*. Finally, Gemini alters structural semantics entirely, changing the definitive Angika phrase *i ten Gulabo chhai* (This is Gulabo) into the demonstrative *yah vah Gulabo hai* (This is that Gulabo).

5 Conclusion

This study addresses the severe digital scarcity of Angika, a marginalized Indo-Aryan language, by introducing the first manually constructed, high-quality parallel corpus of 10,068 sentences. By digitizing authentic regional folklore and meticulously correcting OCR artifacts, The manually

constructed corpus preserves cultural nuances, idiomatic expressions and *Deshaj* (indigenous) vocabulary that are often inadequately represented in synthetic datasets.

To integrate this unseen language into the modern machine translation ecosystem, the NLLB-200 tokenizer was expanded with a dedicated language token for Angika. The ablation study revealed that while phylogenetic warm-initialization from a proximate language (Maithili) provides a safe heuristic, NLLB is robust to token initialization strategies. The critical intervention for extremely low-resource languages is not necessarily how the token is initialized, but rather the architectural registration of a dedicated token combined with targeted fine-tuning.

Furthermore, the comprehensive benchmarking against 8 state-of-the-art LLMs revealed a critical generation-comprehension gap in massively multilingual models. While modern LLMs can comprehend Angika, they fail catastrophically at generating it, defaulting instead to standard Hindi. The lightweight 600M parameter model outperformed LLMs over 100x its size by over +16 BLEU points in the generative (Hindi→Angika) direction. Subsequent human evaluation revealed a “Bilingual Evaluation Paradox,” highlighting the importance of targeted fine-tuning for generating Angika-specific morphology rather than fluent but incorrect standard Hindi.

Ultimately, the methodology demonstrates that the combination of dedicated architectural prompting and authentic, community-driven data collection can successfully bridge the digital divide for extremely low-resource languages.

Acknowledgments

The author deeply thanks Dr. Shobha Kumari, Assistant Professor, Department of Angika at Tilka Manjhi Bhagalpur University (TMBU), Bihar, a native Angika speaker, for her crucial guidance and support in sourcing authentic Angika texts and navigating the linguistic nuances during the translation phase. The author also extends gratitude to Kajal Kumari, linguist and native Angika-Hindi speaker, for her invaluable assistance in the execution of the human evaluation.

References

Anthropic. 2025. *Claude opus 4 & claude sonnet 4*. Technical report, Anthropic.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Sharma, Jason Riesa, Alexis Conneau, and 1 others. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711.

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikui Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the first workshop on neural machine translation*, pages 28–39.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wajah, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Alahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, and 1 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2026. Srcmix: Mixing of related source languages benefits extremely low-resource machine translation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6306–6323.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

- Colin P. Masica. 1991. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge.
- Mistral AI. 2026. [Mistral medium 3.5](#). Mistral model card and changelog.
- OpenAI. 2024. [Hello GPT-4o](#). *OpenAI Blog*.
- Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- 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.
- Ambika Regmi. 2017. A sociolinguistic survey of angika. *Linguistic Survey of Nepal (LinSuN)*.
- Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.
- Ray Smith. 2007. An overview of the tesseract ocr engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2:629–633.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 15894–15939.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Extended Qualitative Error Typology

To provide complete transparency into the morphosyntactic and lexical behaviors of the evaluated architectures, extended translation test set predictions are presented, with all examples rendered in Roman transliteration and italicized.

A.1 Gemini 2.5 Flash: Angika → Hindi

Example 1: Pronominal Number and Gender Resolution Failure

- **Source (Angika):** *ven koy chij ke dekhai, samajhai-būjhai taben soci ke nirṇay lai matur Jhaliya chelai pahārī barsātī nadi nāki khalbal.*
- **Reference (Hindi):** *vah kisī bhī cīz ko dekhtī, samajhī-būjhī tab sockar nirṇay leī, lekin Jhaliya pahārī barsātī nadi kī tarah cañcal thī.*
- **Prediction:** *ve kisī cīj ko dekhte, samajhte-būjhte tab nirṇay lete, lekin vah pahārī barsātī nadi kī tarah khalbalī macāne vālā thā.*
- **Linguistic Analysis:** Angika uses the singular third-person pronoun *ven*. Gemini misinterprets this as the Hindi third-person plural *ve*, altering subject number. Furthermore, because Angika verbs are largely gender-neutral (*dekhai*), Gemini forces a masculine conjugation (*macāne vālā thā*) onto a clearly feminine narrative context, completely omitting the proper noun *Jhaliya*.

Example 2: Semantic Shift of Definitive Particles

- **Source (Angika):** *ī tem Gulabo chai, je khet kem āpno baccā sem bhī jyādā parem dai chai.*
- **Reference (Hindi):** *yah to Gulabo hai, jo khet ko apne bacce se bhī jyādā prem detī hai.*
- **Prediction:** *yah vah Gulabo hai, jo khet ko apne bacce se bhī jyādā pyār kartī hai.*
- **Linguistic Analysis:** The definitive Angika phrase *ī tem Gulabo chai* (meaning “This is Gulabo”) is mistranslated into the demonstrative Hindi structure *yah vah Gulabo hai* (“This is that Gulabo”), distorting structural semantics.

Example 3: Omission of Conjunctive Particles

- **Source (Angika):** *hunī pokhrī mem helī kem ratan tarkī uṭhāy lelkai āro barī premo sem āpnom chātī mem saṭāy lelkai.*
- **Reference (Hindi):** *usne tālāb mem utarkar ratan tarkī uṭhā lī aur bare prem se apnī chātī se lagā lī.*
- **Prediction:** *usne tālāb mem se ratna ko uṭhākar bare prem se apnī chātī se lagā liya.*
- **Linguistic Analysis:** The Angika conjunctive participle *helī kem* (having descended/stepped down) is dropped entirely in the prediction.

Example 4: Grammatical Mood Misalignment

- **Source (Angika):** *jamgal ke rakṣā karbai, gāch-virich nagar vikās ke hor mem bahute kaṭī rahlo chai.*
- **Reference (Hindi):** *jamgal kī rakṣā kareṃge, per-paudhe nagar vikās kī hor mem bahut kaṭ rahe haiṃ.*
- **Prediction:** *jamgal kī rakṣā kareṃ, per-paudhe nagar vikās kī hor mem bahut kaṭ rahe haiṃ.*
- **Linguistic Analysis:** The first-person future declarative *karbai* (“I/we will protect”) is mismapped to the suggestive/imperative Hindi *kareṃ* (“let us protect”).

Example 5: First-Person Pronoun Number Distortion

- **Source (Angika):** *hammem haṃsī ke bol-liyai.*
- **Reference (Hindi):** *maine haṃskar bolā.*
- **Prediction:** *hamne haṃsī mem bolā thā.*
- **Linguistic Analysis:** The singular first-person Angika pronoun *hammem* (I) is mistranslated as the plural Hindi *hamne* (we).

Example 6: Proper Noun Regularization

- **Source (Angika):** *Shukri ke lākh manā kar-lau par vem nai mānlo chelai.*
- **Reference (Hindi):** *Shukri ke lākh manā karne par bhī vah nahīṃ mānā thā.*
- **Prediction:** *shukriyā ke lākh manā karne par bhī vah nahīṃ mānā.*

- **Linguistic Analysis:** The culturally specific proper noun *Shukri* is forcefully regularized into the standard Hindi vocabulary greeting *shukriyā* (thank you).

Example 7: Kinship Entity Hallucination

- **Source (Angika):** *Ramdin kuch nai kahalkai, apno navjāt baccā ārū Lalavatiya kem lai ke gāv se bāhar hoy gelai.*
- **Reference (Hindi):** *Ramdin ne kuch nahīm kahā, apne navjāt bacce aur Lalavatiya ko lekar gāv se bāhar nikal gayā.*
- **Prediction:** *Ramdin ne kuch nahīm kahā, apne navjāt bacce aur patnī ko lekar gāv se bāhar calā gayā.*
- **Linguistic Analysis:** The proper noun *Lalavatiya* is replaced by the generic Hindi kinship noun *patnī* (wife).

Example 8: Emphatic Particle Mistranslation

- **Source (Angika):** *yahī rastā se tem pār hoy cheliyai.*
- **Reference (Hindi):** *isī rāste se to pār hote the.*
- **Prediction:** *isī rāste se tum pār hue the.*
- **Linguistic Analysis:** The emphatic Angika particle *tem* (“indeed/to”) is hallucinated as the second-person Hindi pronoun *tum* (you).

A.2 Gemini 2.5 Flash: Hindi → Angika

Example 9: Maithili Tone Intrusion

- **Source (Hindi):** *Algu aur Jumman bahut pahle hī gāv se svargavāsī ho cuke the.*
- **Reference (Angika):** *Algu-Jumman bahut painhem gāv se svarga vāsī hoy cuklo chai.*
- **Prediction:** *Algu ā Jumman bahut pahine gām sañ svargavāsī bha’ cukal chelai.*
- **Linguistic Analysis:** Exhibits severe Maithili grammatical intrusion, outputting the Maithili ablative *gām sañ* and verbal construction *bha’ cukal*.

Example 10: Lexical Bleed of Maithili Pronouns

- **Source (Hindi):** *mā, tum kuch boltī kyom nahīm ho?*
- **Reference (Angika):** *māy, kahā bolai chiyam kucchu?*

- **Prediction:** *māy, ahā kichū bolai chiyai ki nai?*
- **Linguistic Analysis:** Hallucinates the Maithili second-person honorific *ahā* and quantifier *kichū* rather than Angika equivalents.

Example 11: Verbal Aspect and Loanword Retention

- **Source (Hindi):** *o gāv vāloṃ, sabhī suno! Shanichar pūre gāv ke ṭole-ṭole, galī-galī munādī kartā huā ghūm rahā thā.*
- **Reference (Angika):** *ho gāv vālā sabbhaim sunom, Shanichar sose gāv ṭolā-ṭolā galī-galī ḍholāī denem ghūmī rahlo chai.*
- **Prediction:** *o gāmabāsī sab sunū! Shanichar pūrā gām ke ṭolā-ṭolā, galī-galī munādī karait ghumait rahal chelai.*
- **Linguistic Analysis:** Applies Maithili continuous aspectual morphology (*karait ghumait rahal*) while retaining the Hindi noun *munādī*.

Example 12: Maithili Verbal Roots

- **Source (Hindi):** *Damayanti socī hai - ropāt hone ke bād is kanyā nakṣatra meṃ pānī kā akāl par gayā hai.*
- **Reference (Angika):** *Damayanti socai chai - ropā holā ke bād ī kanyā nakṣatra meṃ pānī kero ṭoyṭā parī gelo chai.*
- **Prediction:** *Damayanti socai chai - ropanī bhelā ke bād ī kanyā nakṣatra meṃ pānī ke akāl pari gelau chai.*
- **Linguistic Analysis:** Utilizes the Maithili verbal root *pari* and retains the standard Hindi vocabulary word *akāl*.

Example 13: Failure on Indigenous Compounds

- **Source (Hindi):** *khetī-mazdūrī ke bal par karz cukāte-cukāte, Sattan kā sharīr-amṡ ṭūṭ gayā thā.*
- **Reference (Angika):** *khetī-majdūrī ke bal par karjā torṭe torṭe, Sattan ke deh-gāt ṭūṭī gelai.*
- **Prediction:** *khetī-majdūrī ke bal par karjā cukaite-cukaite, Sattan ke sharīr-amṡ ṭūṭ gela’ chelai.*

- **Linguistic Analysis:** Retains the Hindi compound *sharīr-amṡ* verbatim rather than predicting the authentic Angika *Deshaj* equivalent *deh-gāt*.

Example 14: Verbatim Stem Retention

- **Source (Hindi):** *sārā badan ekdam pharpharā uṭhā*.
- **Reference (Angika):** *saumse badan ekdam pharpharāy uṭhlai*.
- **Prediction:** *sabhai sharīr ekdam pharpharā uṭhlau*.
- **Linguistic Analysis:** Retains the uninflected Hindi verb stem *pharpharā*.

Example 15: Source Lexical Copying

- **Source (Hindi):** *pūrī sabhā ke logom ne āścharya se donom bhāiyom ko dekhā*.
- **Reference (Angika):** *sose sabhā ro logo acraj se dūnū bhāy ke dekhalkai*.
- **Prediction:** *pūrī sabhā ke lok acraj sañ dunū bhāi kem dekhalkai*.
- **Linguistic Analysis:** Copies the Hindi phrase *pūrī sabhā* literally and inserts the Maithili noun *lok*.

Example 16: Conjunctive Participle Mismatch

- **Source (Hindi):** *usne tālāb meṃ utarkar ratan tarkī uṭhā lī aur baṛe prem se apnī chātī se lagā lī*.
- **Reference (Angika):** *hunī pokhrī meṃ helī kem ratan tarkī uṭhāy lelkai āro baṛī premo sem āpnom chātī meṃ saṭāy lelkai*.
- **Prediction:** *okrā tālāb meṃ utari kay ratan tarkī uṭhāy lelkai ārū baṛ prem sa' apan' chātī sa' lagāy lelkai*.
- **Linguistic Analysis:** Predicts the Maithili conjunctive structure *utari kay* instead of the Angika *helī kem*.

A.3 Fine-Tuned NLLB (600M): Angika → Hindi

Example 17: Transliteration Fallback

- **Source (Angika):** *Damayanti socai chai - ropā holā ke bād ī kanyā nakṣatra meṃ pānī kero ṭoṭāy paṛī gelo chai*.

- **Reference (Hindi):** *Damayanti socī hai - ropā hone ke bād is kanyā nakṣatra meṃ pānī kā akāl paṛ gayā hai*.

- **Prediction:** *Damayanti socī hai - ropā hone ke bād is kanyā nakṣatra meṃ pānī kā ṭoyā paṛ gayā hai*.

- **Linguistic Analysis:** NLLB transliterates Angika regional nouns (*ropā*, *ṭoyā*) directly into the Devanagari output rather than translating them into Hindi semantics.

Example 18: Indigenous Vocabulary Bleed

- **Source (Angika):** *khetī-majdūrī ke bal par karjā torte torte, Sattan ke deh-gāt tūtī gelai*.
- **Reference (Hindi):** *khetī-mazdūrī ke bal par karz cukāte-cukāte, Sattan kā sharīr-amṡ tūt gayā thā*.
- **Prediction:** *khetī-mazdūrī ke bal par karjā torte-torte, Sattan kā deh-gāt tūt gayā*.
- **Linguistic Analysis:** Retains the Angika *Deshaj* compound *deh-gāt* in the Hindi output.

Example 19: Hyphenated Entity Hallucination

- **Source (Angika):** *Algu-Jumman bahut painhem gāv sem svarga vāsī hoy cuklo chai*.
- **Reference (Hindi):** *Algu aur Jumman bahut pahle hī gāv se svargavāsī ho cuke the*.
- **Prediction:** *alag-alag log bahut pahle hī gāv se svargavāsī ho cuke haim*.
- **Linguistic Analysis:** Fails to parse the hyphenated proper noun *Algu-Jumman*, hallucinating the generic phrase *alag-alag log* (different people).

Example 20: Verbal Stem Corruption

- **Source (Angika):** *hunī pokhrī meṃ helī kem ratan tarkī uṭhāy lelkai āro baṛī premo sem āpnom chātī meṃ saṭāy lelkai*.
- **Reference (Hindi):** *usne tālāb meṃ utarkar ratan tarkī uṭhā lī aur baṛe prem se apnī chātī se lagā lī*.
- **Prediction:** *uñhomne pokhrī meṃ helī karke ratan tarkī uṭhā lī aur baṛe pyār se apnī chātī meṃ saṭā liya*.
- **Linguistic Analysis:** Fails to recognize the Angika verb root *helī* (to descend), treating it as an uninflected Hindi noun (*helī karke*).

Example 21: Gender Resolution Interference

- **Source (Angika):** *jo Lakṣmī meharvān rahthiyai te jaj, kiliṭṭar banthiyai matar majbūrī me māṣṭar banī ke māy-bāp ke lāj rākhī lel kai.*
- **Reference (Hindi):** *agar Lakṣmī meharbān hotīm to jaj, kaleṭṭar bantā magar majbūrī meṃ māṣṭar bankar mā-bāp kī lāj rakh lī.*
- **Prediction:** *agar Lakṣmī meharbān rahtī to jaj aur kiliṭṭar bantī lekin majbūrī meṃ māṣṭar bankar mā-bāp kī lāj rakh lī.*
- **Linguistic Analysis:** Because Angika verbs are gender-neutral, NLLB fails gender resolution in Hindi, erroneously predicting the feminine conjugation *bantī* for a masculine narrative subject.

Example 22: Mood Inversion

- **Source (Angika):** *mausī kucchu khāy la' nai deve?"*
- **Reference (Hindi):** *mausī kuch khāne ke liye nahīm dogī?"*
- **Prediction:** *mausī kuch khāne ke liye mat do?"*
- **Linguistic Analysis:** Translates the interrogative future *nai deve* into the prohibitive Hindi imperative *mat do* (do not give).

A.4 Fine-Tuned NLLB (600M): Hindi → Angika

Example 23: Subject-Person Agreement Distortion

- **Source (Hindi):** *jamgal kī rakṣā kareṃge; per-paudhe nagar vikās kī hor meṃ bahut kaṭ rahe haiṃ.*
- **Reference (Angika):** *jamgal ke rakṣā karbai, gāch-ṡirich nagar vikās ke hor meṃ bahute kaṭī rahlo chai.*
- **Prediction:** *jamgal ke rakṣā kartai; gāch-birich nagar vikās ke hor meṃ baṛī kaṭai chai.*
- **Linguistic Analysis:** Maps the first-person plural future *kareṃge* (we will protect) into the third-person Angika future *kartai* (they will protect).

Example 24: Standard Hindi Lexical Copying

- **Source (Hindi):** *o gāv vāloṃ, sabhī suno, Shanichar pūre gāv ke ṭole-ṭole, galī-galī munādī kartā huā ghūm rahā thā.*
- **Reference (Angika):** *ho gāv vālā sabbhaiṃ sunoṃ, Shanichar sose gāv ṭolā-ṭolā galī-galī ḍholāī denem ghūmī rahlo chai.*
- **Prediction:** *he gāvavālā, sab suno! Shanichar saumse gāv ke ṭolā-ṭolā, galī-galī munādī kartem ghūmī rahlo chelai.*
- **Linguistic Analysis:** Retains the Hindi loan-word *munādī* verbatim.

Example 25: Uninflected Stem Bleed

- **Source (Hindi):** *sārā badan ekdam pharpharā uṭhā.*
- **Reference (Angika):** *saumse badan ekdam pharpharāy uṭhlai.*
- **Prediction:** *sose badan ekdam pharpharā uṭhlai.*
- **Linguistic Analysis:** Retains the Hindi verb stem *pharpharā*.

Example 26: Adverbial Retention

- **Source (Hindi):** *main kal baazaar gaya tha.*
- **Reference (Angika):** *hamme kāl bājār gelo cheliye.*
- **Prediction:** *hammeṃ kal bājār gelo cheliyai.*
- **Linguistic Analysis:** Retains the Hindi temporal adverb *kal* instead of the Angika *kāl*.

Example 27: Participial Phrase Copying

- **Source (Hindi):** *dūsre din ke maic meṃ Hindustānī hār ke kagār par pahuñce to Paṃkaj kī dukān par jamī bhīr khel khatm hone se pahle chamṭe gā.*
- **Reference (Angika):** *dūsro din ke maic meṃ Hindustānī ṭīm hār ke kagār par pahuñcī gelai te Paṃkaj ke dukān par jamlo bhīr khel khatam hoy se pahlem chaṭī gelai.*
- **Prediction:** *dosro din ke maic meṃ Hindustānī hār ke kagār par pahuñclai te Paṃkaj ke dukān par jamī bhīr khel khatam hoy ke painhem chaṭkī gelai.*
- **Linguistic Analysis:** Retains the Hindi participial modifier *jamī bhīr* instead of inflecting to Angika *jamlo bhīr*.

Example 28: Possessive Number Mismatch

- **Source (Hindi):** *hamāre to kān kitne bare-bare paṁkhe jaise!*
- **Reference (Angika):** *hamrā sinī keṁ tem kān kattem baraum-baraum paṁkhā raṁ!*
- **Prediction:** *hamro tem kān kattem baro-baro paṁkhā nākhī!*
- **Linguistic Analysis:** Translates the first-person plural Hindi possessive *hamāre* (our) into the Angika singular possessive *hamro* (my).

Example 29: Oblique Nominal Case Omission

- **Source (Hindi):** *Chandrakant ko bete kī bāt jañc gaī.*
- **Reference (Angika):** *Chandrakanto ke betā ke bāt jañcī gelai.*
- **Prediction:** *Chandrakant ke betā ke bāt jañc gelai.*
- **Linguistic Analysis:** Fails to apply Angika's nominal oblique suffix *-o* to the proper noun (*Chandrakanto*).

Example 30: Omission of Inclusive Nominal Morphemes

- **Source (Hindi):** *Sunita bhī sāth jāyegī.*
- **Reference (Angika):** *Sunito sāth jātai.*
- **Prediction:** *Sunita bhī sāth jaitai.*
- **Linguistic Analysis:** Fails to apply Angika's inclusive nominal agglutination (*Sunito*), redundantly retaining the Hindi particle *bhī*.