

When Silver Isn't Enough: Data Quality Effects on Pnar-English Machine Translation

Riya Tekcham

MWire Labs

tekchamriya45@gmail.com

Fitha Asma Sulfeekhar

MWire Labs

fidaasma123@gmail.com

Badal Nyalang

MWire Labs

nyalang@mwirelabs.com

Abstract

Pnar (ISO 639-3: pbv), a low-resource Austroasiatic language spoken in Meghalaya, Northeast India, has no publicly available machine translation systems or parallel corpora benchmarks. This paper investigates data quality effects on Pnar-English machine translation through a systematic ablation across silver-only, gold-only, combined gold+silver, and Khasi-augmented cross-lingual transfer conditions using NLLB-200 and mBART-50. Results show that data quality consistently outweighs data quantity for this language pair, with the combined gold+silver NLLB-200 model achieving the best automatic metric performance (BLEU 30.39 en→pnar, 26.02 pnar→en on the gold test set). Cross-lingual transfer from the related language Khasi leads to BLEU degradation. Human evaluation was not conducted for the Khasi-transfer outputs limiting our ability to determine whether semantic content was actually preserved despite the decline in BLEU. Human evaluation on 50 sentences by two native Pnar speakers yields mean adequacy of 4.25 and fluency of 4.53 out of 5, with strong inter-annotator agreement (quadratic weighted $\kappa = 0.70$ adequacy, 0.91 fluency). Data and models will be released upon acceptance under CC-BY 4.0.

1 Introduction

Pnar (also known as Jaintia; ISO 639-3: pbv) is an Austroasiatic language of the Khasian branch spoken primarily in the Jaintia Hills of Meghalaya, Northeast India. Typologically, Pnar is verb-final, tonal, and morphologically agglutinative, with a rich system of noun class markers (Ring, 2015). It is closely related to Khasi and shares significant lexical and structural overlap, as established by lexicostatistical analysis (Nagaraja et al., 2013). Despite its speaker population, Pnar remains absent from mainstream NLP benchmarks and MT systems, and has received no coverage in major multilingual model pretraining pipelines.

Recent multilingual NLP work for Northeast India has begun addressing the broader resource gap. Nyalang (2026) introduced masked language models for nine Northeast Indian languages including Pnar, but MT for Pnar has not been explored. Related languages such as Khasi and Mizo have received attention in MT (Thabab and Purkayastha, 2021) and resource development (Ghosh et al., 2025), and Khasi was included in the WMT 2024 low-resource Indic translation shared task (Pakray et al., 2024). Pnar was not part of any of these efforts. The availability of massively multilingual models like NLLB-200 (NLLB Team et al., 2022) and mBART-50 (Tang et al., 2021) has opened pathways for low-resource MT without requiring large in-domain parallel corpora. However, the quality and composition of fine-tuning data remains an open question for extremely low-resource settings where gold supervision is scarce and silver data may not reflect authentic language use. This paper asks: for a language like Pnar, does more data always help, or does data quality dominate?

Contributions of this paper are as follows: (1) we present systematic ablation experiments across five data conditions for Pnar-English MT using two pretrained multilingual models; (2) we show that combining gold and silver data consistently outperforms silver-only and gold-only training; (3) we find that cross-lingual transfer from Khasi, tested specifically to determine whether typological similarity benefits Pnar, degrades surface-level quality while preserving semantic content; and (4) we provide two-annotator human evaluation of the best-performing system with inter-annotator agreement.

2 Background and Related Work

2.1 Low-Resource MT for Northeast Indian Languages

Low-resource MT for Indian languages has advanced through initiatives such as Samanantar

(Ramesh et al., 2022), IndicTrans2 (Gala et al., 2023), and the NLLB project (NLLB Team et al., 2022). The WMT 2024 low-resource Indic shared task (Pakray et al., 2024) covered four Northeast Indian language pairs including English-Khasi and English-Mizo. Early transformer-based MT for Khasi was explored by Thabab and Purkayastha (2021), and resource development for Mizo and Khasi across NLP tasks was undertaken by Ghosh et al. (2025). Pnar has not appeared in any of these initiatives, and no prior parallel corpus or MT system for the language has been made publicly available.

2.2 Silver Data and Back-Translation

The use of back-translated or synthetic silver data in MT is established practice (Sennrich et al., 2016). For extremely low-resource languages, silver data quality is harder to control and may fail to represent authentic in-domain language use. Our work quantifies this directly for Pnar using web-scale silver data sourced from FineTranslations (Penedo et al., 2026), a large-scale dataset produced by translating non-English web content into English using the Gemma 3 27B model (Gemma Team, 2025).

2.3 Cross-Lingual Transfer

Pnar and Khasi share the Khasian branch of Austroasiatic with significant lexical overlap (Nagaraja et al., 2013). NLLB-200 includes Khasi in its pretraining data. Our cross-lingual transfer experiments are designed specifically to test whether Khasi benefits Pnar, not the reverse: all transfer conditions fine-tune a Pnar-trained model on Khasi data, and evaluation is on the Pnar gold test set only.

3 Data Preparation

3.1 Parallel Corpus Construction

The English source sentences for the gold dataset are drawn from the Tatoeba corpus (Tiedemann, 2020), a community-contributed multilingual sentence collection. Pnar translations were produced by two paid native Pnar translators: one a linguistics postgraduate and the other completing undergraduate studies in a related field. All translations were human-verified. The Silver corpus was obtained from the Pnar-English portion of the FineTranslations dataset hosted on Hugging Face (Penedo et al., 2026). FineTranslations is constructed from FineWeb2, a multilingual web cor-

pus, with the original web documents translated into English. The source metadata in the dataset identifies Wyrta, a regional Pnar news media outlet, as the source of the original Pnar text. The corresponding English translations were synthetically generated using Gemma 27B (Gemma Team, 2025), as documented in the dataset resources. The corpus therefore primarily represents the news domain. A manual overview of the corpus indicated that the Pnar source sentences were generally well-formed, while the synthetic English translations exhibited some variation in quality. The Pnar-English pairs were provided as aligned document-level translations. Given the complete absence of prior parallel data for Pnar, the gold corpus of 4,826 training pairs represents a meaningful baseline resource sufficient to establish initial translation quality benchmarks for the language. We found no exact overlap between the Gold and Silver corpora, and no duplicate sentence pairs were present in the final Silver as well as gold corpus.

3.2 Sentence Segmentation

We apply language-specific preprocessing for sentence segmentation and tokenization across the dataset. For English text, we use the Natural Language Toolkit (NLTK) sentence tokenizer to segment raw text into sentence boundaries. For Pnar, we adopt a rule-based segmentation approach using custom regular expressions due to the absence of a dedicated NLP toolkit for the language which posed several challenges during data preparation. A major challenge involved handling abbreviations and honorifics such as “Mr.”, “Dr.”, etc. Even though English text was processed using NLTK, it still failed in certain cases present in the dataset. An example from the silver corpus includes phrase like “Dr. Feeling Roy Suchinag”, NLTK tokenizes as [’Dr.’, ’Feeling Roy Suchinag’]. NLTK correctly identifies “Dr.” as an abbreviation, but because “Feeling” is also a valid English word, it misinterprets the structure and leads to incorrect segmentation or boundary issues within named entities. In such cases, the toolkit fails to preserve the full name as a single entity. Because of this, we had to develop additional custom regex-based rules not only for Pnar, but also for handling edge cases in English text to ensure correct sentence segmentation and alignment. We also removed cases where English and Pnar produced a mismatched number of sentences within the same paragraph. In such cases, the alignment was considered unreliable, so

Split	Silver	Gold	Combined
Train	48,519	4,826	53,345
Validation	6,012	604	604
Test	5,937	604	604

Table 1: Pnar-English dataset statistics. Validation and test sets for the combined condition use gold data only.

the entire paragraph was dropped to avoid introducing noise into the dataset.

3.3 Data Filtering

After segmentation, we applied length based filtering and length ratio based filtering to remove noisy and misaligned sentence pairs. Sentence pairs with lengths outside the range of 2 to 150 tokens were removed. In addition, we applied a length ratio filter between source and target sentences, keeping only pairs with a ratio between 0.21 and 4.5. This helped eliminate extreme mismatches.

3.4 Dataset Statistics

Table 1 summarises the dataset splits used across all experiments. To ensure a consistent and reliable evaluation across training configurations, we use the Gold test set as the common evaluation benchmark. Since the Silver corpus consists of automatically generated translations and may contain translation noise, evaluating on the manually curated Gold test set provides a higher-quality and consistent reference for comparing the different systems. The validation splits were retained as part of the dataset partitioning but were not used for checkpoint selection or model selection, as all models were trained for a fixed number of epochs.

4 Models and Experimental Setup

Two pretrained multilingual MT models are fine-tuned:

- **NLLB-200** (NLLB Team et al., 2022): 600M parameter distilled model (facebook/nllb-200-distilled-600M) with explicit Khasi coverage in pretraining. Custom language tokens pnar_Latn (ID 256204) and kha_Latn (ID 256205) were added to the tokenizer and the model embedding matrix resized accordingly.
- **mBART-50** (Tang et al., 2021): Many-to-many multilingual denoising model (facebook/mbart-large-50-

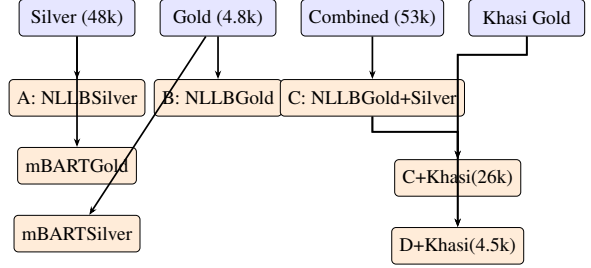


Figure 1: Experimental design. Arrows indicate fine-tuning data sources. Models C+Khasi and D+Khasi are initialised from Model C and further fine-tuned on Khasi gold data to test directional cross-lingual transfer into Pnar.

many-to-many-mmt). Combined data experiments were not pursued for mBART-50 given its weaker baseline performance on gold-only conditions; this remains future work.

Both translation directions are trained simultaneously by concatenating original and direction-flipped datasets. All models are fine-tuned using the AdamW optimiser with learning rate 5e-5, batch size 8, maximum sequence length 128 tokens, and mixed precision training (fp16) for 5 epochs on an A40 GPU. Seven experimental conditions are evaluated in total. Figure 1 shows the experimental design and data flow.

5 Results

5.1 Automatic Evaluation

All models are evaluated using BLEU (Papineni et al., 2002), ChrF (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020), computed using the Hugging Face evaluate library. As COMET reference models are not trained on Pnar data, scores are reported as a complementary neural signal and should not be treated as language-specific quality assessments. NLLB-200 results appear in Table 2 and mBART-50 results in Table 3. To assess the statistical reliability of differences between systems, we perform paired bootstrap resampling with 1,000 samples using SacreBLEU. The bootstrap procedure provides confidence intervals for the BLEU scores and a paired significance test between the compared systems. In particular, we compare the Gold-only and Combined systems on the common Gold test set.

5.2 Human Evaluation

Human evaluation was conducted on 50 sentences randomly selected from the gold test set, translated

Model	Test Set	BLEU	ChrF	TER	COM.
<i>EN → PNAR</i>					
A (Silver)	G	9.44	29.65	82.34	.588
A (Silver)	S	34.70	55.68	55.98	.701
B (Gold)	G	28.11	50.95	55.17	.687
C (G+S)	G	30.39	53.29	52.74	.693
C (G+S)	S	34.69	55.72	56.05	.701
C+Khasi (26k)	G	3.94	30.11	88.23	.701
D+Khasi (4.5k)	G	4.83	31.50	93.80	.619
<i>PNAR → EN</i>					
A (Silver)	G	7.16	27.58	100.01	.539
A (Silver)	S	34.60	54.68	59.49	.704
B (Gold)	G	25.42	45.79	61.72	.685
C (G+S)	G	26.02	46.75	60.37	.700
C (G+S)	S	34.55	54.68	59.49	.705
C+Khasi (26k)	G	16.11	38.21	79.31	.634
D+Khasi (4.5k)	G	19.57	40.74	72.10	.661

Table 2: NLLB-200 results. Test Set: G = Gold, S = Silver. G+S = Gold+Silver. COM. = COMET.

Model	Dir.	BLEU	ChrF	TER	COM.
Gold	en→pnar	25.80	49.16	58.84	.673
Silver	en→pnar	26.80	49.58	73.85	.652
Gold	pnar→en	9.19	27.76	94.73	.450
Silver	pnar→en	19.23	44.91	107.16	.577

Table 3: mBART-50 results on the gold test set. COM. = COMET.

by Model C (en→pnar). Two native Pnar speakers independently rated each output on a 1–5 scale for adequacy (degree of meaning transfer) and fluency (naturalness in Pnar). Annotator 1 is a native speaker with research experience in the language; Annotator 2 is a journalist with over 20 years of professional experience in regional language media. Annotators were not informed of each other’s ratings.

Inter-annotator agreement was substantial: quadratic weighted Cohen’s $\kappa = 0.70$ for adequacy and 0.91 for fluency, with Spearman correlations of 0.789 and 0.853 (both $p < 0.001$). Table 5 presents individual and aggregate scores.

Combined mean adequacy was 4.25 and fluency 4.53, indicating generally natural Pnar output with reliable meaning transfer. Qualitative analysis of outputs is provided in Appendix A.

6 Analysis

Data quality outweighs quantity. Model A, trained on 48,519 silver pairs, scores only 9.44 BLEU on the gold test set for en→pnar, compared to 28.11 for Model B trained on 4,826 gold pairs. This is a tenfold difference in training data yielding a threefold improvement in the opposite direc-

Direction	Gold CI	Combined CI	p
EN→PNAR	± 1.59	± 1.63	0.005
PNAR→EN	± 1.95	± 1.98	0.178

Table 4: Paired bootstrap significance test p -values for Gold-only and Combined systems on the Gold test set. CI denotes the bootstrap confidence-interval half-width based on 1,000 resamples.

Annotator	Adequacy		Fluency	
	Mean	SD	Mean	SD
Annotator 1	4.37	0.85	4.59	0.64
Annotator 2	4.13	0.73	4.47	0.60
Combined	4.25	–	4.53	–
Quad. weighted κ : 0.70 (adq.), 0.91 (flu.)				

Table 5: Human evaluation results for Model C (en→pnar), 50 sentences, 1–5 scale.

tion. Silver data distribution does not reflect authentic Pnar usage and fails under clean evaluation conditions. Combining gold and silver (Model C) yields the best results across all metrics, suggesting the two sources are complementary: silver provides broad lexical coverage while gold anchors the model to authentic Pnar.

NLLB-200 outperforms mBART-50. NLLB-200 consistently outperforms mBART-50 across all conditions. The gap is largest in pnar→en: mBART Gold scores 9.19 BLEU versus NLLB-200 Gold’s 25.42. This likely reflects NLLB-200’s explicit inclusion of Khasi in pretraining, providing stronger priors for the Khasian language family. For mBART-50, silver training marginally outperforms gold in en→pnar (26.80 vs. 25.80 BLEU), a reversal of the NLLB-200 pattern that may reflect mBART-50’s weaker Khasian priors making it more dependent on data volume.

Khasi transfer degrades BLEU but preserves semantics. The Khasi transfer experiments were designed specifically to test whether typological similarity between Khasi and Pnar benefits translation into Pnar. Sequential fine-tuning on Khasi gold data causes BLEU to drop sharply to 3.94 (en→pnar, 26k condition), while COMET remains at 0.701. However, since the COMET model has not been validated for Pnar, the stable COMET score should be interpreted only as a complementary neural signal and cannot by itself be taken as evidence of semantic preservation. The observed degradation may reflect lexical interference or other effects arising from sequential fine-tuning between closely

related languages. Furthermore, human evaluation was not conducted for the Khasi-transfer outputs, limiting our ability to determine whether semantic content was actually preserved despite the decline in BLEU. Neither the 26k nor the 4.5k condition improves over Model C on any metric.

TER above 100. TER scores exceeding 100 for Model A (pnar→en gold) and mBART Silver (pnar→en) indicate outputs requiring more edit operations than the reference length, consistent with hallucination or off-target generation.

Statistical reliability. For English-to-Pnar translation, the paired bootstrap test yields a statistically significant difference ($p = 0.005$). The bootstrap confidence intervals for the Gold-only and Combined systems are ± 1.59 and ± 1.63 , respectively. In contrast, for Pnar-to-English translation, the difference is not statistically significant ($p = 0.178$), with confidence intervals of ± 1.95 and ± 1.98 for the Gold-only and Combined systems, respectively. These results provide statistical support for the improvement observed in English-to-Pnar translation, whereas the improvement in the reverse direction should be interpreted cautiously.

7 Conclusion

This paper establishes MT baselines for Pnar-English (pbv), showing that data quality plays an important role in this low-resource language pair, although the observed gains may also be influenced by domain differences between the Gold and Silver corpora. NLLB-200 fine-tuned on combined Gold and Silver data achieves the strongest results. Cross-lingual transfer from Khasi, tested specifically in the direction that could benefit Pnar, does not improve translation quality at the data scales tested and results in a substantial decline in BLEU. Human evaluation by two native Pnar speakers corroborates that the best system produces generally adequate and fluent translations. Data and models will be released upon acceptance under CC-BY 4.0.

Limitations

The Gold dataset is limited in size (4,826 training pairs), and expanding it remains an important direction for improving translation quality. The Gold and Silver corpora also come from different source domains. Although the training and test sets within each corpus follow their respective data distributions, the Gold training data are

more closely aligned with the Gold test set than the Silver data. Therefore, the better performance of the Gold data may be influenced by both data quality and domain similarity, and cannot be attributed entirely to data quality. COMET scores should be interpreted with caution because the underlying reference models have not been validated for Pnar. In particular, the stable COMET score observed after Khasi fine-tuning does not by itself provide sufficient evidence that semantic content was preserved. Human evaluation was also not conducted for the Khasi-transfer outputs, limiting our ability to assess semantic preservation in this setting. The Khasi transfer experiments use sequential fine-tuning. Other approaches, such as Khasi pre-adaptation followed by Pnar fine-tuning, joint or multi-task training with language tags, data mixing strategies, and parameter-efficient methods such as adapters or LoRA, may produce different results and are left for future work. Combined-data experiments were not conducted for mBART-50 because of its weaker gold-only baseline and are also left for future work. Finally, all experiments use a fixed hyperparameter configuration without tuning and are based on a single training run per condition. Multi-seed experiments would provide a more comprehensive assessment of training variability and are left for future work. Our experiments do not systematically investigate different Gold–Silver ratios. Future work will evaluate progressively larger Silver subsets and Gold oversampling strategies to identify effective mixtures and better characterize the quality–quantity trade-off.

Acknowledgement

We thank MWire Labs for providing computational resources and support for this work. We are grateful to the MWire Labs team for their assistance with data creation and annotation. We also thank our native Pnar speaker annotators for their careful evaluation of model outputs.

References

- Jay Gala, Raj Dabre, Riyaz Ahmad Bhat, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages](#).
- Gemma Team. 2025. [Gemma 3](#).
- Soumyadip Ghosh, Henry Lalsiam, Dorothy Marbaniang, Gracious Mary Temsen, Rahul Mishra,

- and Parameswari Krishnamurthy. 2025. [Towards resource-rich Mizo and Khasi in NLP: Resource development, synthetic data generation and model building](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 228–239, Vienna, Austria. Association for Computational Linguistics.
- K. S. Nagaraja, Paul Sidwell, and Simon Greenhill. 2013. A lexicostatistical study of the Khasian languages: Khasi, Pnar, Lyngngam, and War. *Mon-Khmer Studies*, 42:1–11.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). Preprint, arXiv:2207.04672.
- Badal Nyalang. 2026. [NE-BERT: A multilingual language model for nine Northeast Indian languages](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. [Findings of WMT 2024 shared task on low-resource Indic languages translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.
- 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Amir Hossein Karagaran, and Leandro von Werra. 2026. FineTranslations. <https://huggingface.co/datasets/HuggingFaceFW/finetranslations>.
- 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, Lisbon, Portugal. Association for Computational Linguistics.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Hiram Ring. 2015. *A Grammar of Pnar*. Ph.D. thesis, Nanyang Technological University, Singapore. PhD thesis.
- 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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. [Multilingual translation from denoising pre-training](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- N. Donald Jefferson Thabab and Bipul Syam Purkayastha. 2021. [Low resource neural machine translation from English to Khasi: A transformer-based approach](#). In *Proceedings of the International Conference on Computing and Communication Systems*, volume 170 of *Lecture Notes in Networks and Systems*, pages 3–13. Springer, Singapore.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

A Qualitative Analysis of Model C Outputs

Table 6 presents representative en→pnar translation examples from Model C across three quality categories: high quality (both adequacy and fluency rated 5.0 by both annotators), low quality (both dimensions rated 3.0), and fluency-adequacy mismatch (high fluency, low adequacy).

High-quality outputs tend to involve short, concrete sentences with direct Pnar equivalents. Errors

Source (EN)	Model C Output (Pnar)	Adq.	Flu.
<i>High quality</i>			
We must take care of our planet, the earth.	Toh u sumar i ya ka pyrthai yong i, ka pyrthai.	5.0	5.0
Answer at once when spoken to.	Jubab syndet manda klam ya ki.	5.0	5.0
<i>Low quality</i>			
Lucky at cards, unlucky in love.	Ka jingmyntoi ha ki kard, ka jingma kawa sepsngi.	3.0	3.0
In high altitudes, people find it hard to breathe.	Ha ki thaw wa jrong, ki bru chem jingeh ki wow dih snam.	3.0	3.0
<i>Fluent but inadequate</i>			
If you have questions, ask away.	Lada em jingkylli phi, kylli cha lait.	2.5	4.25

Table 6: Representative Model C (en→pnar) outputs with mean human evaluation scores (1–5 scale, averaged across two annotators). Adq. = Adequacy, Flu. = Fluency.

in lower-quality outputs include idiomatic expressions that lack direct Pnar translations (e.g., “lucky at cards, unlucky in love”) and domain-specific vocabulary (e.g., altitude-related terminology). The fluency-adequacy mismatch case illustrates a pattern where the model produces grammatically natural Pnar but shifts the pragmatic meaning — “ask away” rendered as “kylli cha lait” (literally “ask and go”) is fluent but loses the permissive sense of the original.