# Machine Translation for Bhojpuri, Magahi Maithili using low resource monolingual and bilingual corpus

**Stage 1:** Setup and Custom Tokenization (Hindi-English & Bhojpuri).
**Stage 2:** Fine-tuning IndicTrans.
**Stage 3:** Fine-tuning NLLB-200
**Stage 4:** Optimization for Magahi

# Stage 1

- We first validated our training pipeline using Hindi-English data.
- The base IndicTrans model did not support Bhojpuri well.
- We created a custom tokenizer for Bhojpuri using monolingual data.
- This allowed better handling of Bhojpuri text.

# Stage 3

- **Task:** Translate between Hindi and Bhojpuri.
- **Model:** Meta's NLLB-200 (600M parameters).
- **Method:** Full fine-tuning of the model parameters.
- **Result:** BLEU score improved from **7.63** (baseline) to **26.31**.

# Stage 4

- **Target:** Hindi $\leftrightarrow$ Magahi translation.
- **Problem:** Very little parallel data available (only ~800 sentence pairs).
- **Solution:** We needed advanced techniques to prevent overfitting.

# Stage 4

- **Synthetic Data:** We used back-translation to create ~4,000 synthetic training pairs.
- **Efficiency:** We used Low-Rank Adaptation (LoRA) to fine-tune fewer parameters.
- **Optimization:** Loaded the model in 8-bit mode to save memory.

# Results

- The optimized model performed very well on the test set.
- **sacreBLEU Score:** 37.24
- **chrf Score:** 65.10
- **Conclusion:** Synthetic data and LoRA are effective for low-resource languages.