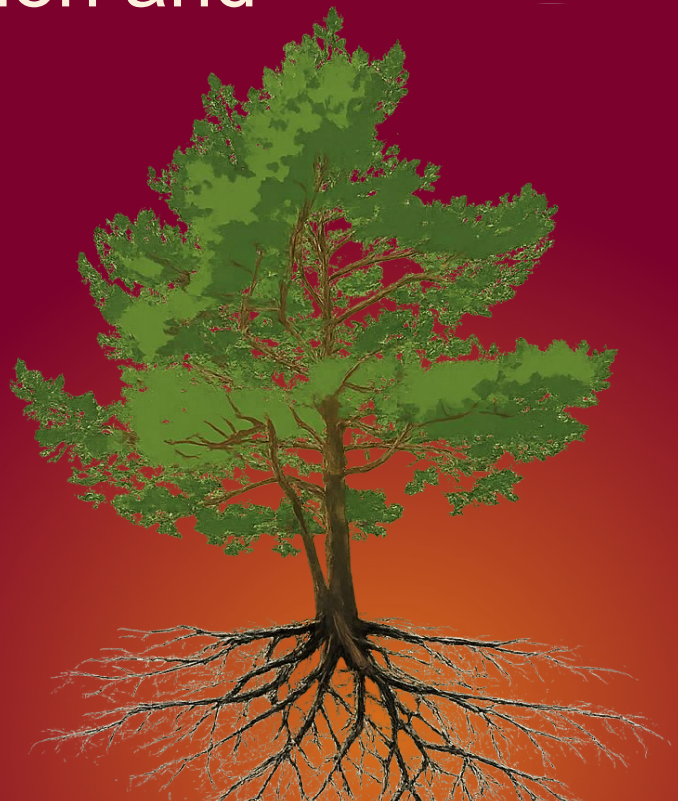

KWÄN DÉK'ÁN' DO
TO KEEP THE FIRE BURNING

AI Translation Supporting Northern Tutchone Language Revitalization and Strengthening Cultural Roots

Jordan Alexander

Saad Hasan

25 February 2025



Translation is constrained by data scarcity, model coverage, and governance requirements

Decline of Living Fluent Speakers + Data Scarcity

Living fluent speakers are rapidly disappearing, and there is a lack of recorded language.

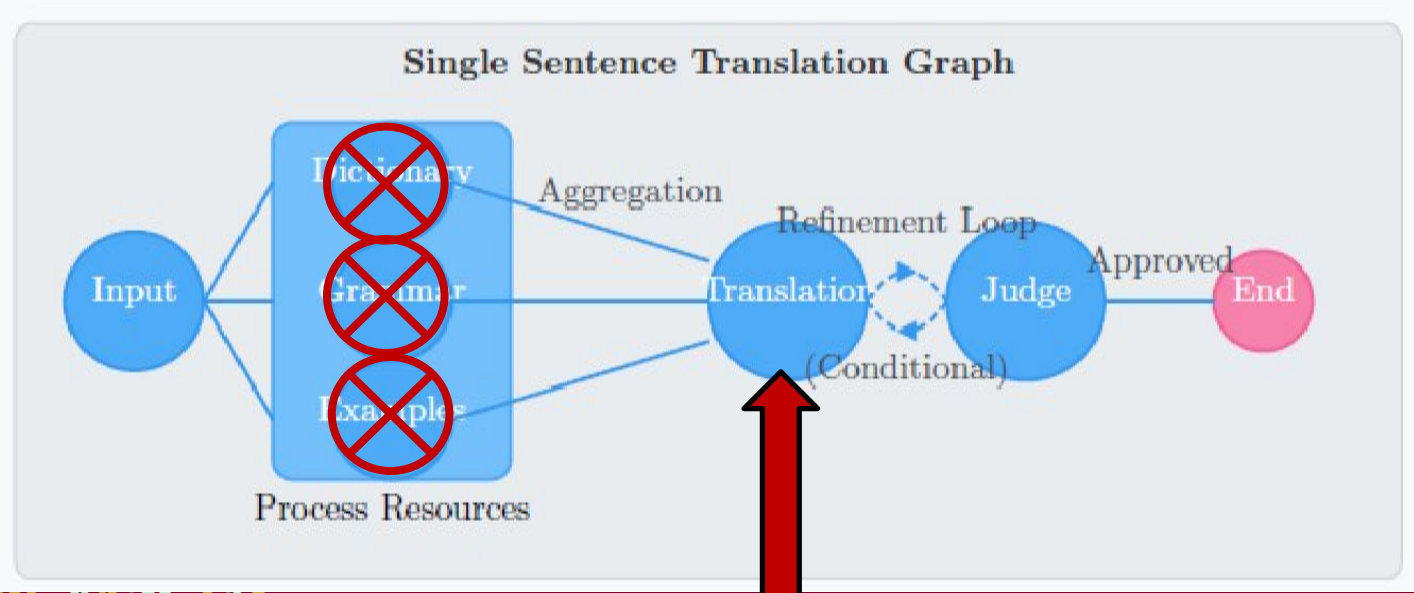
No coverage in Major LLM's or Translation Models

Major large language models lack coverage of the language.

Data Governance & Safety Constraints

Data governance constraints are required for respectful data handling and safety.

A prompt-based approach was useful, but it could not scale into a reliable translator



Open source model

Fine-tuned language models can make translation feasible for very low-resource languages

ByT5

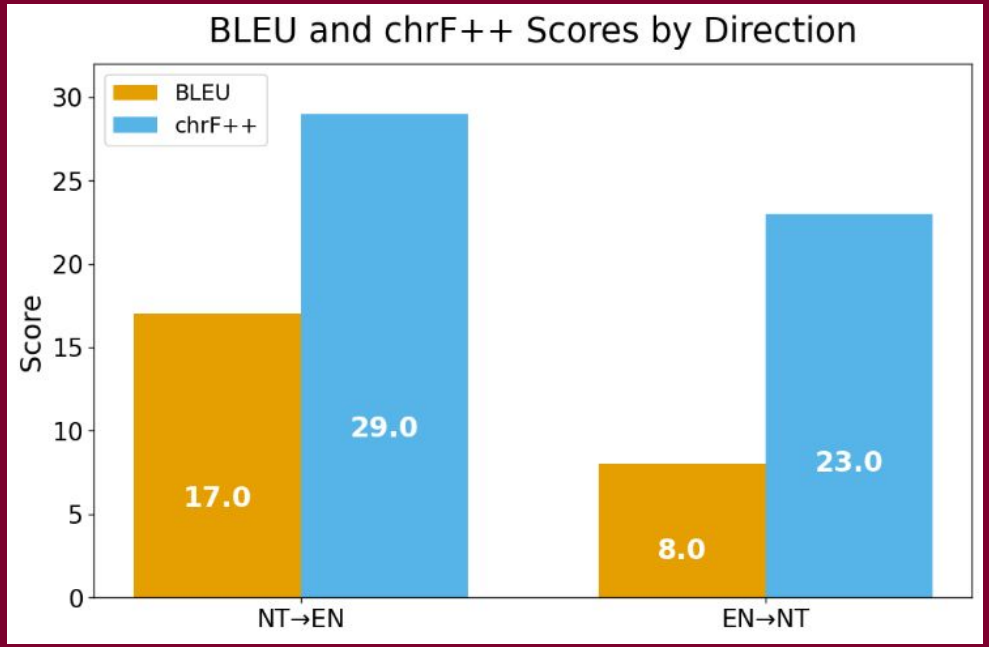
- Byte-level tokenization
- Robust to any language
- Fine-tuneable
- Train/Run locally on consumer hardware

mT5

- Subword tokenization
- Strong cross-lingual priors
- Fine-tuneable
- Train/Run locally on consumer hardware

Current ByT5-small (LoRA) baseline & results

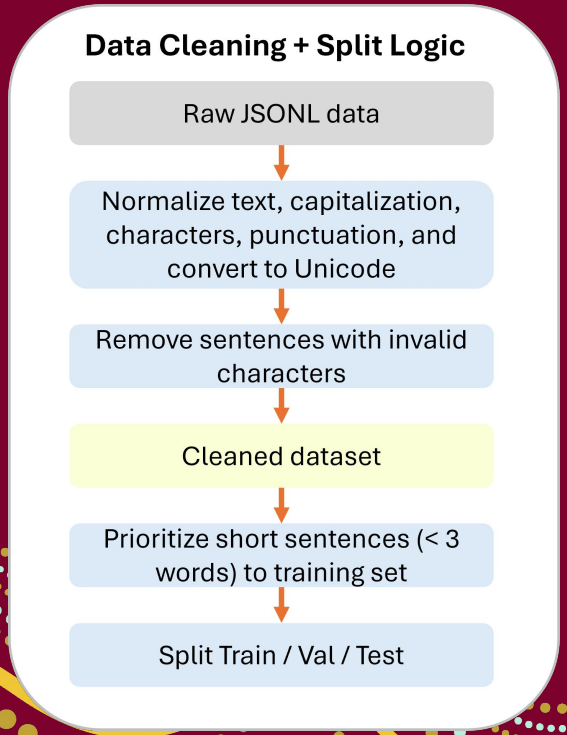
- **~1.8k sentence pairs** post-cleaning (from ~3000 original)
- 3% perfect translations (NT □ EN, val set)

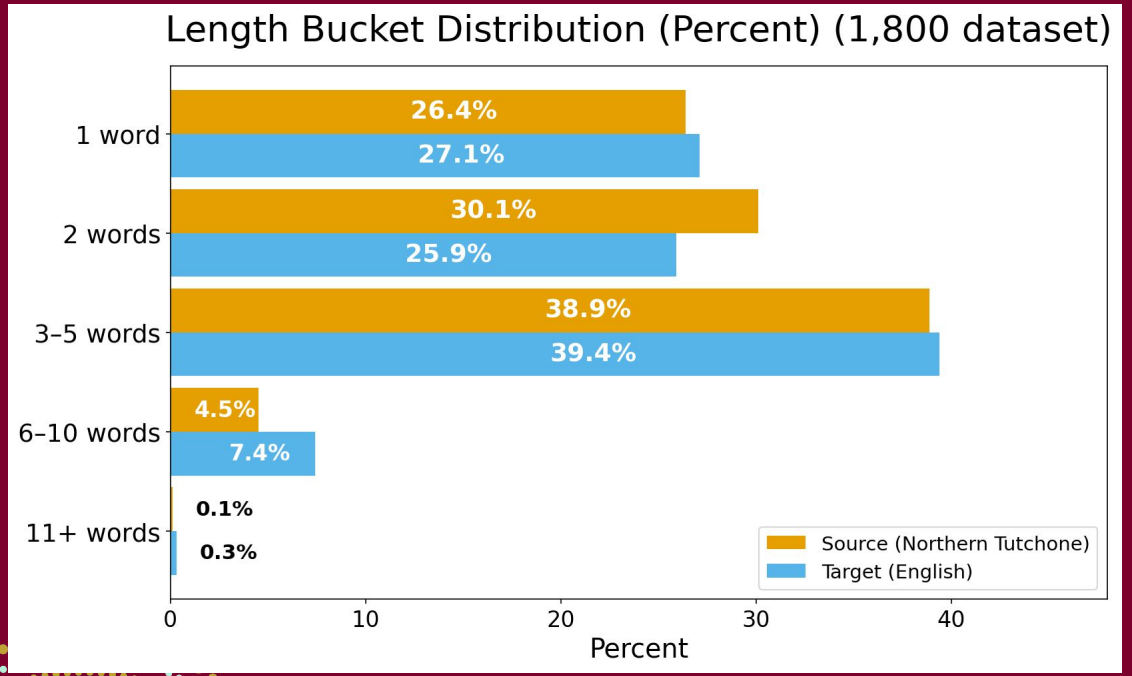


Current ByT5-small (LoRA) baseline & results

	NT Sentence	Reference/Actual Translation	Predicted Translation
1	Tedhaw nenínthyáw.	Warm up the soup .	She is wearing a warm jacket .
2	Dànálát tíhìnchín !	Take our boat out of the water !	Take your boat out of the water !
3	Dek'ą́ą łądézhe.	The boy is going hunting.	The boy is going hunting.

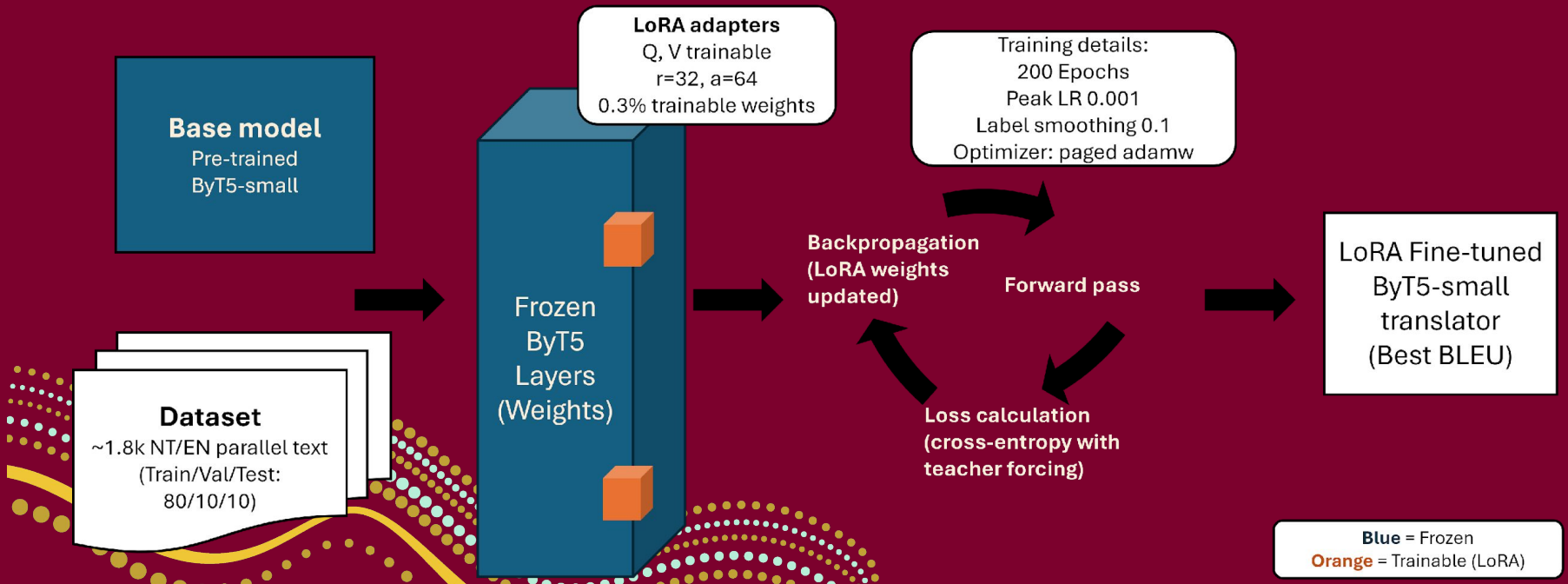
Data preparation, preprocessing and evaluation pipeline





Architecture and training process

1. Data Input → 2. Solution Design (LoRA fine-tuning) → 3. Training Process (Fine-Tuning Loop) → 4. Output & Evaluation



Ethical considerations

Data Governance & Community Focus

- Consent & stewardship
- OCAP adherence

Accuracy & Safety

- Hallucination risks
- Safe usage

Environmental Footprint

- Small, local model
- Very low energy usage

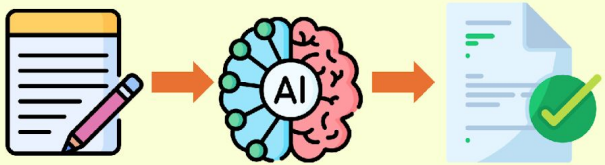
Where the project is going

Baseline Improvement
- Train on new data
- Parameter tuning
- mT5 experiments

Data Growth
- Data augmentation
- Data generation

Community Integration
- Humans-in-the-loop

Language Learning Tool



MUSSI CHO