THOTH Validation Report

Generated: 2025-12-06 04:05:18

Evaluation Dataset: FLORES-200 (facebook/flores200)

Sample Size: 200 sentences per language

Executive Summary

This validation evaluated THOTH's two translation engines against the FLORES+ benchmark.

- **Total translations**: 6,000
- **Total errors**: 6
- **Evaluation time**: 156.8 minutes

Score Comparison Table

All Languages (chrF Score)

Language	NLLB chrF	NLLB BLEU	Argos chrF	Argos BLEU	Winner
Russian	60.3	34.1	62.5	35.9	Argos
Ukrainian	62.1	36.9	50.7	23.4	NLLB
Lithuanian	57.6	30.5	N/A	N/A	NLLB*
Latvian	58.7	31.8	N/A	N/A	NLLB*
Estonian	61.9	35.6	N/A	N/A	NLLB*
Serbian	65.6	41.4	N/A	N/A	NLLB*
Bulgarian	66.1	40.6	N/A	N/A	NLLB*
Polish	56.6	28.7	56.0	26.5	Tie
Czech	63.9	38.5	62.9	36.2	Tie
Greek	59.5	34.0	58.1	30.8	Tie
Chinese (Simplified)	56.1	27.6	54.6	24.5	Tie
Japanese	53.0	23.3	46.4	16.6	NLLB
Korean	55.4	26.6	45.6	15.1	NLLB
Arabic	63.8	40.0	57.2	30.9	NLLB
French	69.4	46.1	69.2	45.7	Tie
German	67.0	42.8	64.2	37.3	NLLB
Spanish	59.9	31.0	53.9	22.8	NLLB

Language	NLLB chrF	NLLB BLEU	Argos chrF	Argos BLEU	Winner
Norwegian Bokmål	63.9	39.6	N/A	N/A	NLLB*

** Argos not available for this language*

Aggregate Statistics

NLLB-200

- **Languages evaluated**: 18
- **Mean chrF**: 61.15 (std: 4.47)
- **Mean BLEU**: 34.95 (std: 6.27)
- **chrF range**: 53.0 - 69.4
- **Total translation time**: 8784.0s

Argos Translate

- **Languages evaluated**: 12
- **Mean chrF**: 56.77 (std: 7.15)
- **Mean BLEU**: 28.81 (std: 9.03)
- **chrF range**: 45.6 - 69.2
- **Total translation time**: 612.5s

Tier Analysis

Tier 1 - Critical Languages

- **NLLB mean chrF**: 60.12 (5 languages)
- **Argos mean chrF**: 56.60 (2 languages)

Tier 2 - Important Languages

- **NLLB mean chrF**: 62.34 (5 languages)
- **Argos mean chrF**: 58.99 (3 languages)

Tier 3 - Coverage Languages

- **NLLB mean chrF**: 61.06 (8 languages)
- **Argos mean chrF**: 55.87 (7 languages)

Recommendations

Based on this evaluation:

1. **Default Engine**: Use NLLB-200 as the primary engine due to:
   - Broader language coverage (200 languages vs ~45 for Argos)
   - Consistent quality across language families
   - Support for all critical Tier 1 languages

2. **When to consider Argos**:
    - Languages where Argos outperforms: Russian (+2.2 chrF)
3. **Language-specific notes**:

Output Files

File	Description
`summary_report.md`	This executive summary
`scores_by_language.csv`	Per-language chrF and BLEU scores
`scores_by_engine.csv`	Aggregate statistics per engine
`detailed_results.csv`	Sample sentence-level results
`engine_recommendations.md`	Engine recommendations by language