Project Report

Dataset Selection

- **Source:** English-French Wiki dataset from Opus.
- **Link** : Opus
- **Initial Sentence Pairs:** 818,302.

Filtering Process

1. **Duplicates Removed:** 803,704 pairs.
2. **Sentences ≤ 200 Words:** 801,392 pairs.
3. **Length Ratio ≤ 1.5:** 691,348 pairs.
4. **Non-printable Characters Removed:** 691,222 pairs (126 removed).

Preprocessing Pipeline

- **Character Cleaning:** Removed non-printable/control characters from both source and target.
- **Normalization:** Applied NFKC normalization and reduced multiple spaces to single spaces**.**
- **Symbol Removal:** Eliminated unwanted symbols while retaining essential punctuation using regex.

COMET Scoring and Sampling

- **Scoring:** Calculated COMET (wmt20-comet-qe-da) scores for 50% of the data (345,611 pairs).
- **Language Detection (Source):**

Language	Count	Language	Count
en	316,161	ca	971
fr	11,872	pt	818
de	4,039	nl	668
it	2,339	id	628
unknown	1,357	sv	528
es	1,166	ro	499
tl	1,057	af	481

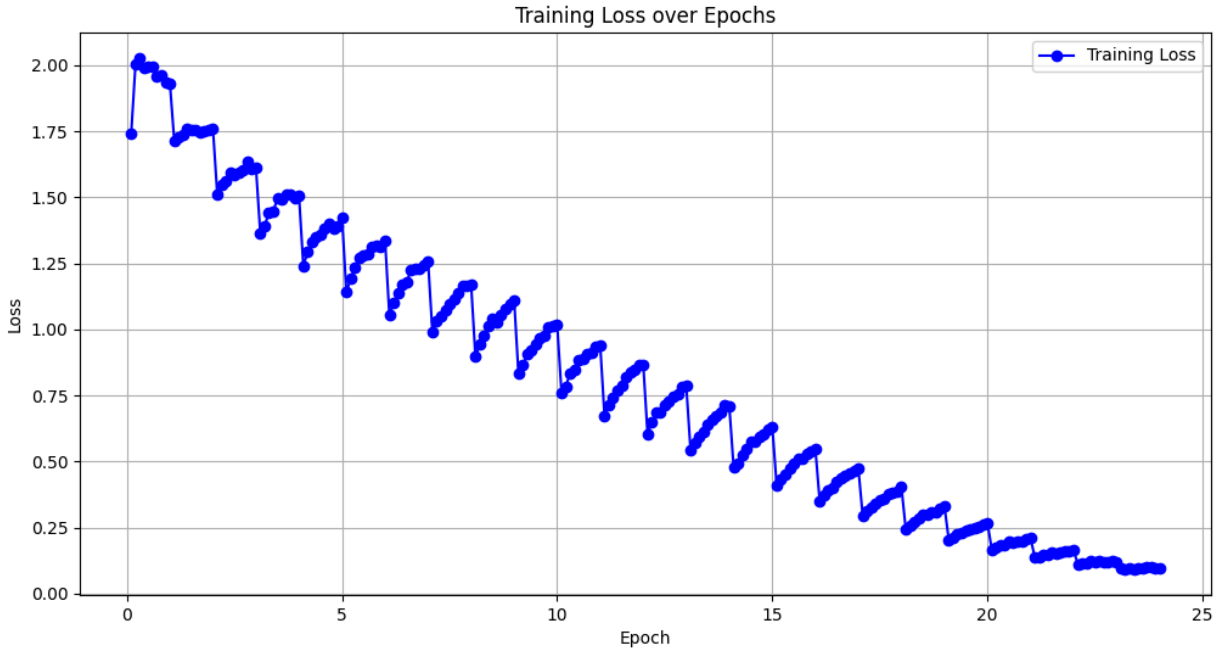
- **Language Detection (Target):**

Language	Count	Language	Count
fr	298,086	nl	184
en	14,621	ro	145
de	611	pt	142
ca	549	unknown	98
it	540	id	82
es	460		

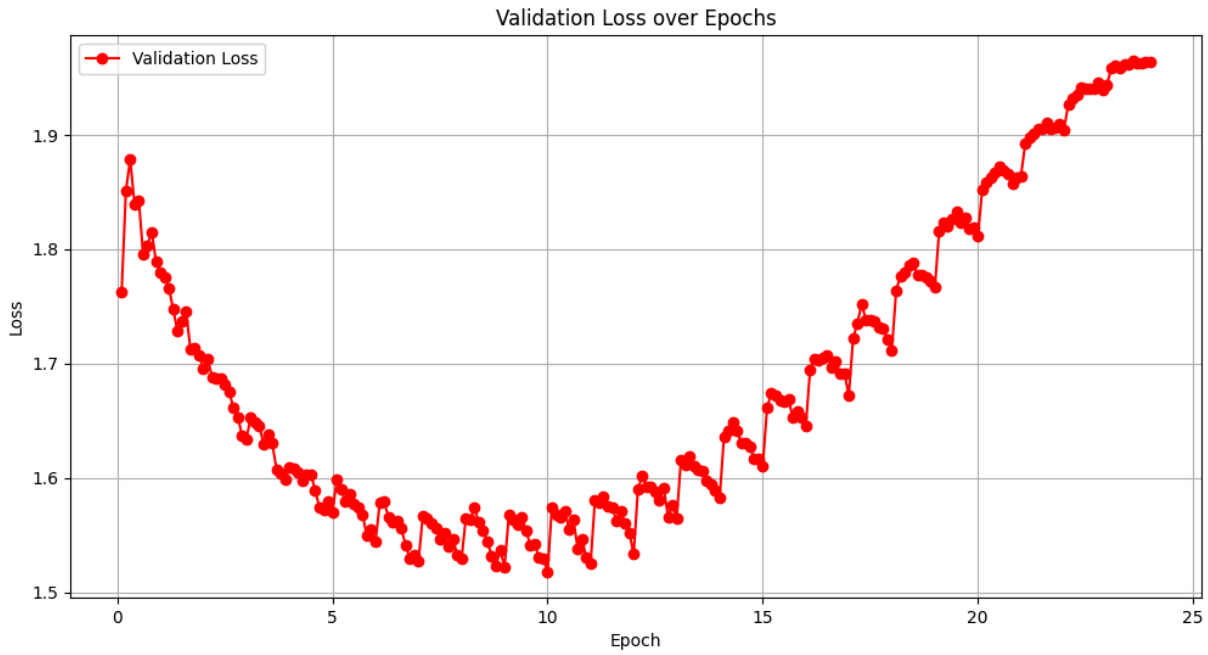
- **Filtered Rows:** 298,086 pairs after removing sentences from wrong languages (en-fr).
- **Sampling:** Randomly selected 130K pairs, split into 100K (train), 15K (validation), and 15K (test).

Model Fine-Tuning

- **Model Used:** Helsinki-NLP/opus-mt-en-fr from Hugging Face.
- **Training Metrics:**
    - **Training Loss**



- **Validation Loss:**



So the model overfitted after 10 Epochs. Hence we used the checkpoint from Epoch 10 for Evaluation.

Performance Evaluation

- **Baseline:**
  - **SacreBLEU:** 42.20
  - **chrF++:** 65.99
  - **COMET (Unbabel/wmt20-comet-qe-da):** 0.395
- **Fine-Tuned Model:**
  - **SacreBLEU:** 43.55
  - **chrF++:** 69.16
  - **COMET (Unbabel/wmt20-comet-qe-da):** 0.566

Recommendations for Improved Performance

1. **Multi-Directional MT Models:** Utilize many-to-many translation models to enhance performance across various language directions by leveraging shared learning.
2. **Advanced Training Techniques:** Explore Reinforcement Learning from Human Feedback (RLHF) methods like KTO, which have shown superior results compared to Supervised Fine-Tuning (SFT) in previous studies.