UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

Can fine-tuning multilingual sentence embeddings improve the accuracy of Russian–Turkmen bitext retrieval and semantic similarity tasks?

View to “Köpetdag” mountains. Capital city Ashgabat, Turkmenistan.

Jelaleddin Sultanov | Data Science Seminar | Prof. Dr. Michael Granitzer, Dr. Jelena Mitrovic

Motivation & Context

Why this project?

- Bureaucracy in Turkmenistan is handled in **Turkmen & Russian** due to historical legacy.
- There are **no robust MT models** for Turkmen—language resources are minimal.
- Scanned/unscraped books exist in Turkmen and Russian, offering untapped parallel text.

Data & Resources

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

1. Tatoeba: open-source parallel dataset 160K sentence pairs (not clean & topic biased).
2. Uzbek e-library: high-school textbooks verified by Uzbekistan's Ministry of Education, published in **Russian, Turkmen, Uzbek, Tajik, Kazakh.**
3. **Turkmen News website - news that published in Russian and Turkmen languages**

	Tatoeba	Uzbek e-library: high school books	Turkmen News website
Data Collection	HuggingFace/Github	Books available at: https://uzedu.online/	Web Scraping: https://www.turkmenportal.com/ https://orient.tm
Data cleaning	No data cleaning, used as it is	Capitalization, blank space removal. Fixing errors from results of sentence embedding mappings	Capitalization, blank space removal. Fixing errors from results of sentence embedding mappings
Sentence pairs	160K	30K	20K
Data Quality	1. Heavily biased to one topic: Jehovah's Witnesses 2. Semantic errors	Consistent grammar, syntax, errors	Semantic errors

Table 1: Overview of parallel corpus data

Data & Resources

Some description about STS17 dataset

Data Curation Workflow - Uzbek high-school e-books

1. Scrape and extract text from Uzbek-hosted Turkmen and Russian textbooks, Turkmen news website
2. Generate embeddings via pretrained embedding models
3. Map sentence pairs using cosine similarity using generated embeddings
4. Manually Clean & label dataset with **Label Studio**

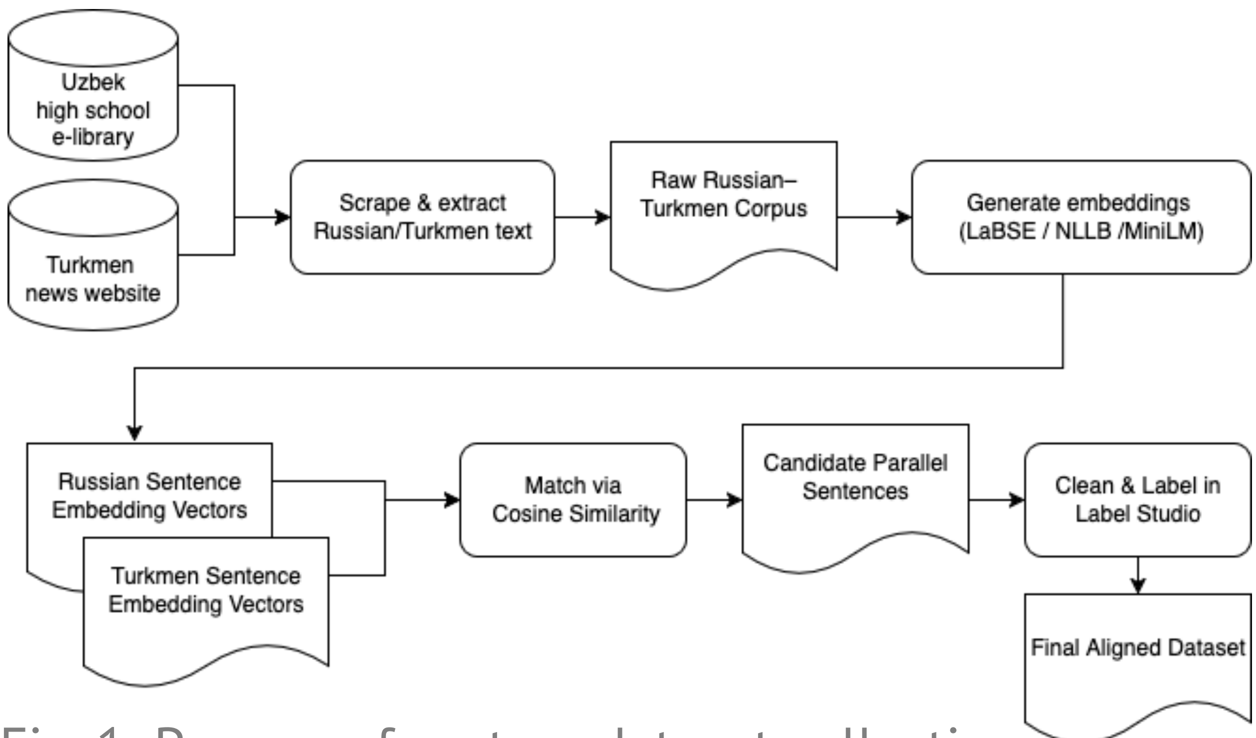
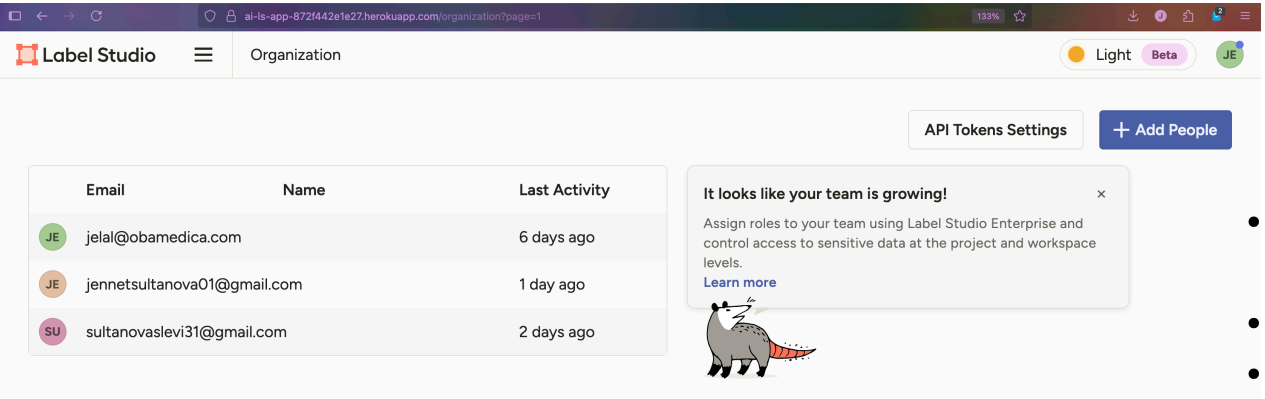


Fig. 1: Process of custom dataset collection

Label Studio Setup for data annoatation

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics



- People working on annotation: 3 (More people are to join)
- Hosted on heroku.com EU servers
- Around 50K sentence pairs collected in total

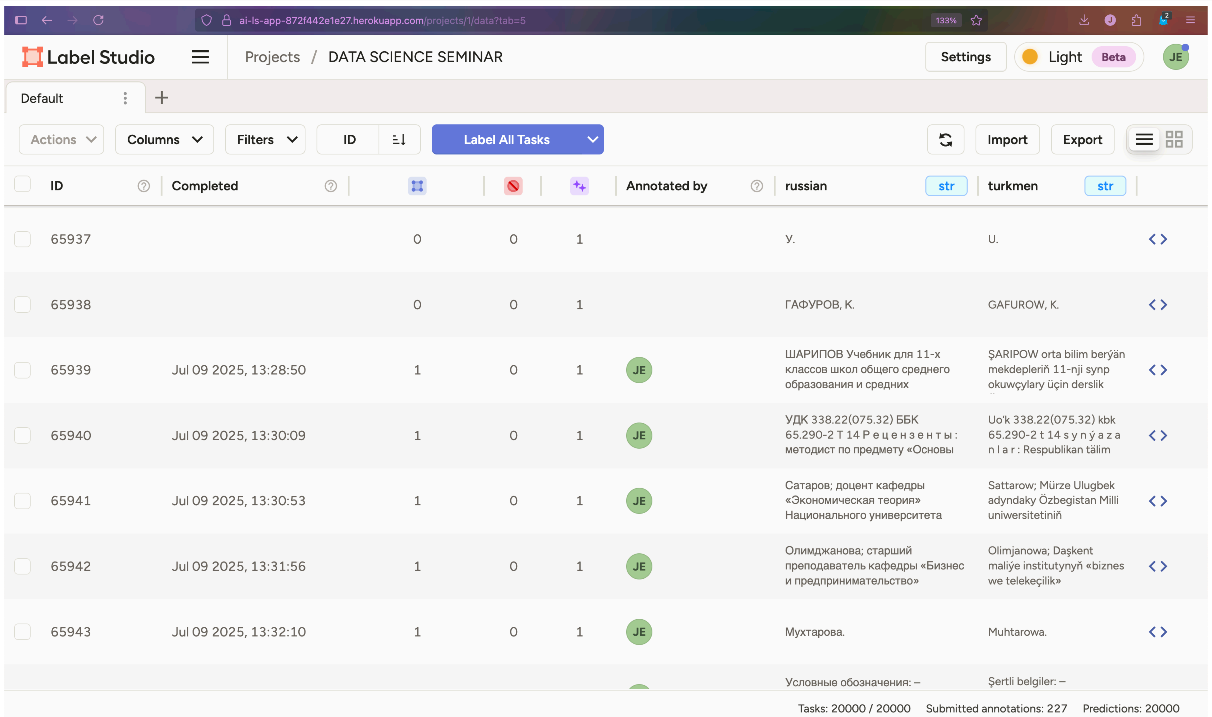


Fig.2 Label Studio Dashboard

Embedding Models Overview

- **LaBSE** – optimized for cross-lingual bitext retrieval
- **NLLB** – Seq2Seq MT model; embeddings aligned with translation output
- **MiniLM** – lightweight, efficient, and excels in semantic similarity

Related Work

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

LaBSE (Feng et al., 2020)

- BERT-based dual-encoder model trained with translation ranking loss on parallel corpora.
- Strong performance in multilingual retrieval and similarity tasks.
- https://arxiv.org/abs/2007.01852

NLLB (Team et al., 2022)

- Encoder-decoder model trained on 200+ languages using FLORES and No Language Left Behind pipeline.
- Optimized for low-resource MT and zero-shot capabilities.
- https://arxiv.org/abs/2207.04672

MiniLM (Wang et al., 2020)

- Lightweight student model distilled from multilingual transformers.
- Used widely in retrieval and sentence similarity tasks.
- https://arxiv.org/abs/2012.15828

Model Overview: NLLB (No Language Left Behind)

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

What is NLLB?

- Developed by Meta AI in 2022
- A **Seq2Seq (encoder-decoder)** model trained for **multilingual machine translation**
- Supports over **200 languages**, including low-resource ones like **Turkmen**

Key Features:

- Based on the **Transformer architecture**
- Uses a **shared encoder** across languages
- Capable of generating translation and embeddings from the same model

Why I Used It:

- My long-term goal is to build a Turkmen MT system
- NLLB gives **encoder embeddings** that can be directly used for sentence comparison
- Enables **fine-tuning for translation quality** and also bitext alignment

Model Overview: LaBSE (Language-Agnostic BERT Sentence Embedding)

University of Passau
Faculty of Computer Science and Mathematics

What is LaBSE?

- Developed by Google Research
- A **dual-encoder model** fine-tuned with **translation ranking loss** on parallel corpora
- Embeds 109+ languages into a **shared vector space**

Key Features:

- Based on **BERT architecture**
- Trained to **maximize similarity between translations**
- Excellent for **bitext retrieval and sentence alignment**

Why I Used It:

- Specially designed for **cross-lingual sentence matching**
- Ideal for low-resource alignment tasks like **Russian–Turkmen**
- Fine-tuning with **LoRA adapters** improved performance on my dataset

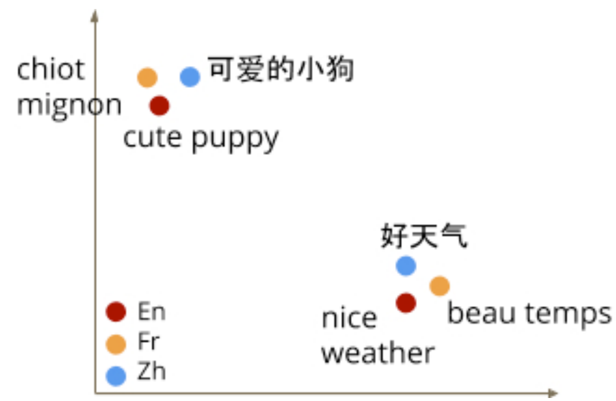


Fig. 3: LaBSE
Credits: [LaBSE official webpage]

Model Overview: MiniLM (paraphrase-multilingual-MiniLM-L12-v2)

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

What is MiniLM?

- A **lightweight transformer** developed by Microsoft & HuggingFace
- Trained using **knowledge distillation** from larger multilingual models
- Supports 50+ languages in **compact architecture**

Key Features:

- Only **12 transformer layers**, making it fast and memory-efficient
- Delivers **competitive semantic similarity results**
- Used widely in **semantic search and retrieval**

Why I Used It:

- Small model with **surprisingly high performance** in STS (Pearson r = 0.7893)
- Useful as a **baseline for semantic similarity**
- Helpful for **scaling embeddings on large datasets**

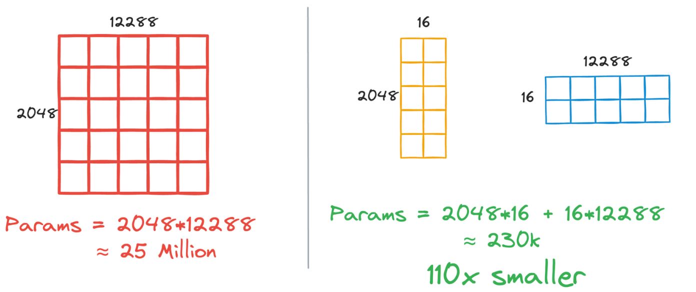
NLLB & LoRA: Efficient Fine-Tuning Technique

What is LoRA?

- **LoRA (Low-Rank Adaptation)** is a method for fine-tuning large models efficiently by inserting **small trainable layers** into frozen pretrained models
- Instead of updating all parameters, it learns **low-rank updates (matrices A & B)** in selected attention layers

Why LoRA?

- Saves GPU memory
- Faster training
- Works well with multilingual transformers like LaBSE or NLLB



How I Used It:

- Fine-tuned **LaBSE** using LoRA adapters on my custom rus–tuk dataset
- Only adapter layers were trained – base model remained untouched
- Result: Bitext retrieval improved → P@1 = **1.000**, MRR = **1.000**

Image credits: https://www.dailydoseofds.com/implementing-lora-from-scratch-for-fine-tuning-llms/

Quick overview of LoRA technique

UNIVERSITY OF PASSAU
Faculty of Computer Science and Mathematics

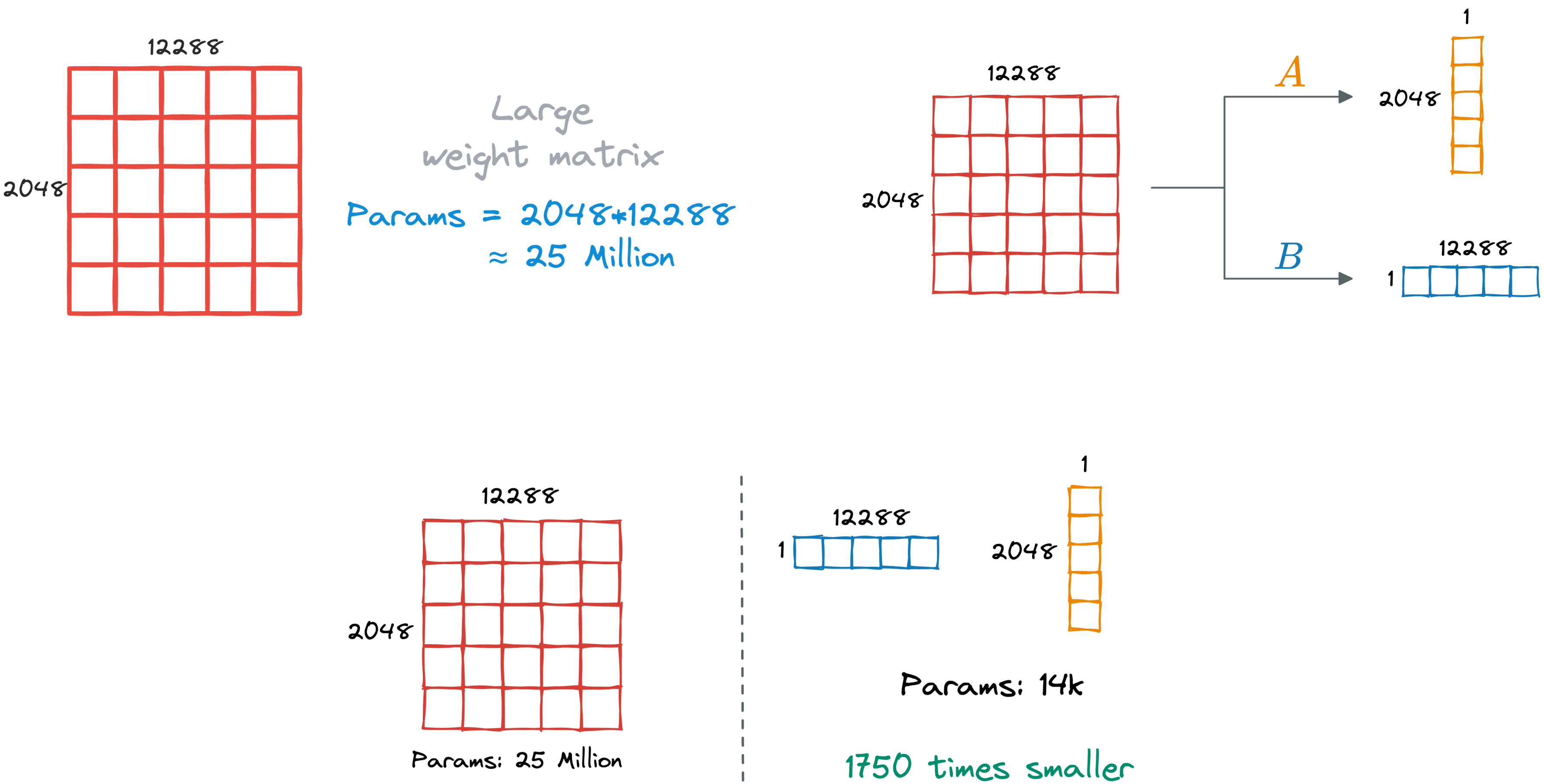


Image credits:
https://www.dailydoseofds.com/implementing-lora-from-scratch-for-fine-tuning-llms/

Hypothesis:

Fine-tuning LaBSE, NLLB, and MiniLM on Russian–Turkmen data will improve P@1, MRR, and Pearson r over pretrained baselines.

Methodology:
Data Preparation

- **Dataset**:

  a) custom **Turkmen-Russian parallel corpus** was used for training and evaluation.

  b) open-source Tatoeba dataset

  c) STS17 dataset

- **Preprocessing**: Tatoeba dataset text was cleaned (removing unwanted characters and extra spaces)

- **Tokenization**: The **LaBSE**, **NLLB**, and **MiniLM** models were tokenized using the respective language codes (**tuk_Latn** for Turkmen and **rus_Cyrl** for Russian).

Methodology:
Fine-Tuning Approach

Faculty of Computer Science and Mathematics

- **Models Used**:
    - **NLLB-200 (Distilled)**: Pretrained by Meta and fine-tuned for multilingual translation tasks using the LoRA technique.
    - **LaBSE (Language-Agnostic BERT Sentence Embedding)**: Optimized for cross-lingual sentence matching.
    - **MiniLM**: A lightweight model for efficient semantic similarity tasks.
- **LoRA Configuration**:
    - Focused on **attention layers** (q_proj and v_proj) for efficient fine-tuning.
    - **Hyperparameters**: r=16, lora_alpha=32, lora_dropout=0.05, ensuring efficient use of memory and faster training.
- **Training Settings**:
    - **Google Colab Pro** - NVIDIA **A100 (40GB) GPU**
    - **Epochs**: 5
    - **Batch Size**: 8 per device
    - **Learning Rate**: 1e-5
    - **Evaluation Metric**: **BLEU** for model selection, based on performance on the validation set.

Methodology:

Metrics & Evaluation:

- **Bitext Retrieval**: Evaluated using **Precision@1 (P@1)** and **Mean Reciprocal Rank (MRR)** for sentence alignment.
- **Semantic Similarity**: Evaluated using **Pearson's correlation coefficient (r)** on the **STS17 Russian dataset** to assess how well the embeddings capture semantic similarity.
- **Other Metrics**: **BLEU**, **chrF**, and **TER** (for machine translation) were also calculated to compare the fine-tuned models against the pretrained versions.

Evaluation Tasks & Metrics

1. **Bitext Retrieval:**

- **P@1 (Precision at 1)**: Measures the accuracy of the top-ranked retrieval; checks if the correct translation is first in the list.
- **MRR (Mean Reciprocal Rank)**: Measures the average rank of the first correct translation; higher values indicate quicker retrieval of the correct answer.

2. **Semantic Textual Similarity (STS):**

- **Pearson's r**: Measures the linear correlation between predicted and actual similarity scores; higher values indicate better alignment with human judgments.

3. **Machine Translation:**

- **BLEU**: Measures n-gram precision in the translation; higher scores indicate better translation quality.
- **chrF**: Measures character-level similarity, useful for languages with rich morphology; higher scores indicate better translation.
- **TER**: Measures the number of edits needed to match the reference translation; lower scores indicate fewer edits and better quality.

Results: Bitext Retrieval

Model	P@1	MRR
NLLB (pretrained)	1.0	1.0
NLLB (fine-tuned)	1.0	1.0
LaBSE (pretrained)	0.8889	0.9444
LaBSE (fine-tuned)	1.0	1.0
MiniLM (pre-trained)	0.5556	0.7011

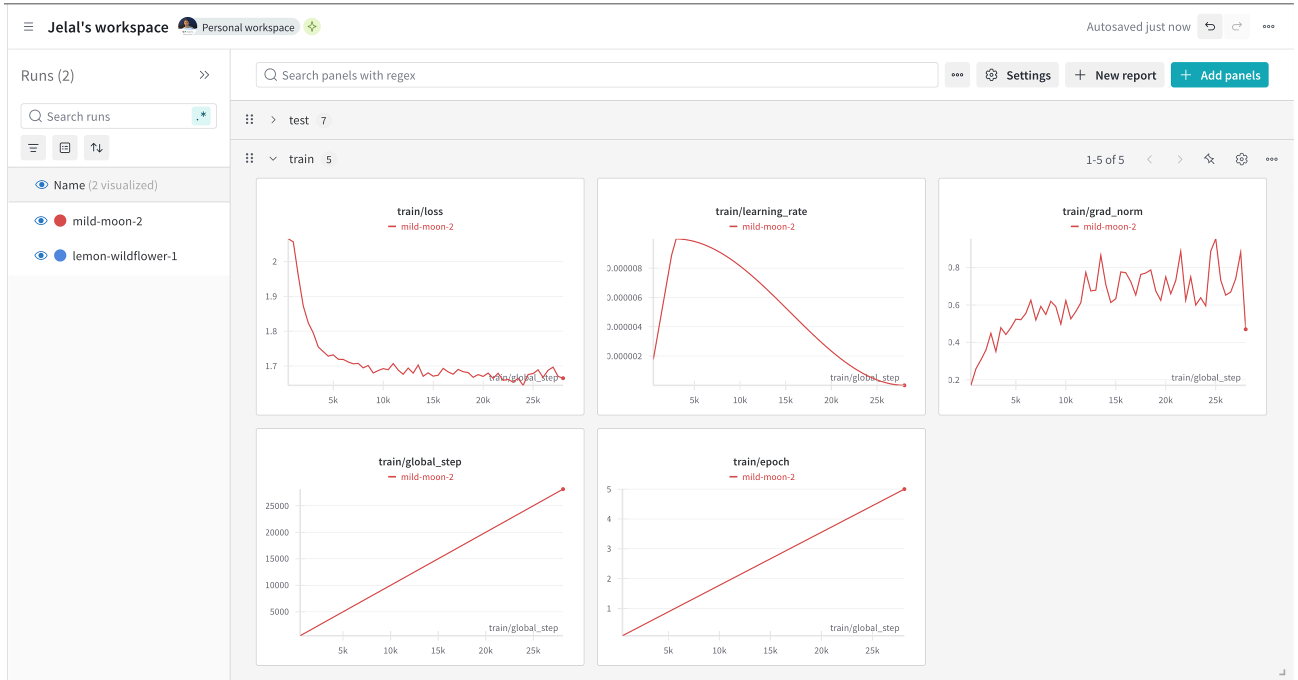
More on NLLB finetuning on next slide

Results: STS Evaluation

University of Passau
Faculty of Computer Science and Mathematics

Model	P@1
NLLB (pretrained)	0.6996
NLLB (fine-tuned)	0.6994
LaBSE (pretrained)	0.7357
LaBSE (fine-tuned)	0.6715
MiniLM (pre-trained)	0.7893

Results: NLLB MT Fine-Tuning



W&B Dashboard - Logs of NLLB Training

Evaluating Translation for Russian —> Turkmen

	Fine-tuned	Pretrained
BLEU	16.64	16.44
chrF	40.52	39.17
TER	76.73	80.08

Evaluating Translation for Turkmen —> Russian

	Fine-tuned	Pretrained
BLEU	16.82	16.24
chrF	35.97	37.49
TER	90.38	85.97

Results Analysis

Bitext Retrieval:

- All fine-tuned models (NLLB, LaBSE) achieved **P@1 = 1.0** and **MRR = 1.0** → perfect top-1 alignment.
- MiniLM (pretrained) showed lower retrieval (P@1 = 0.56), confirming its limitations for alignment tasks.
- Fine-tuning LaBSE significantly boosted retrieval (from 0.89 → 1.0), but may risk overfitting to aligned pairs.

STS Semantic Similarity:

- Best performer: **MiniLM (pretrained)** with **Pearson r = 0.7893** — strongest in capturing sentence meaning.
- LaBSE fine-tuning slightly worsened performance (from 0.7357 → 0.6715), suggesting loss of general semantic strength.
- NLLB showed **no gain from fine-tuning** (r ~0.699), indicating it's better kept frozen for STS.

Thank You!