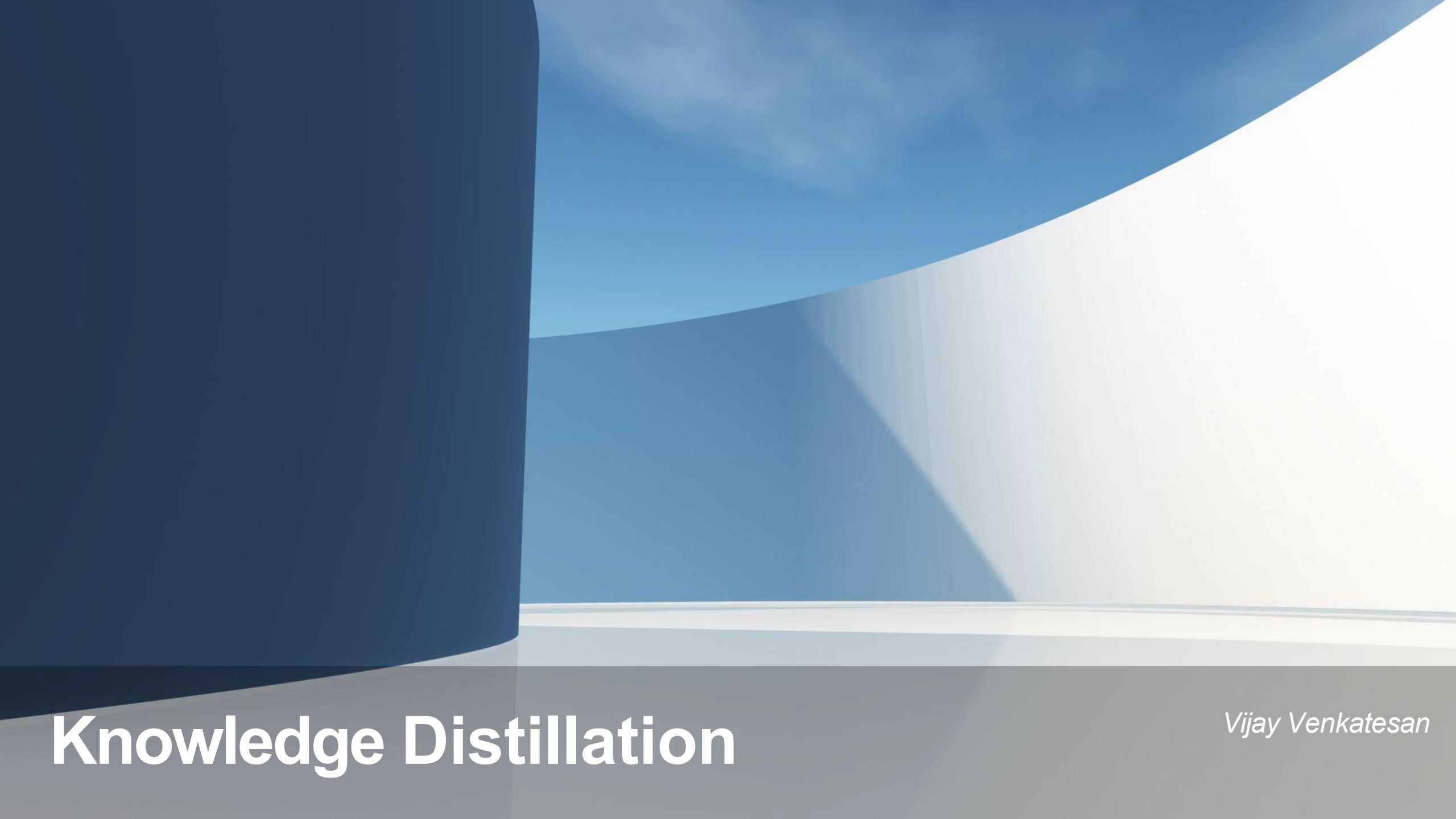


# Knowledge Distillation

*Vijay Venkatesan*

# A Quick Review

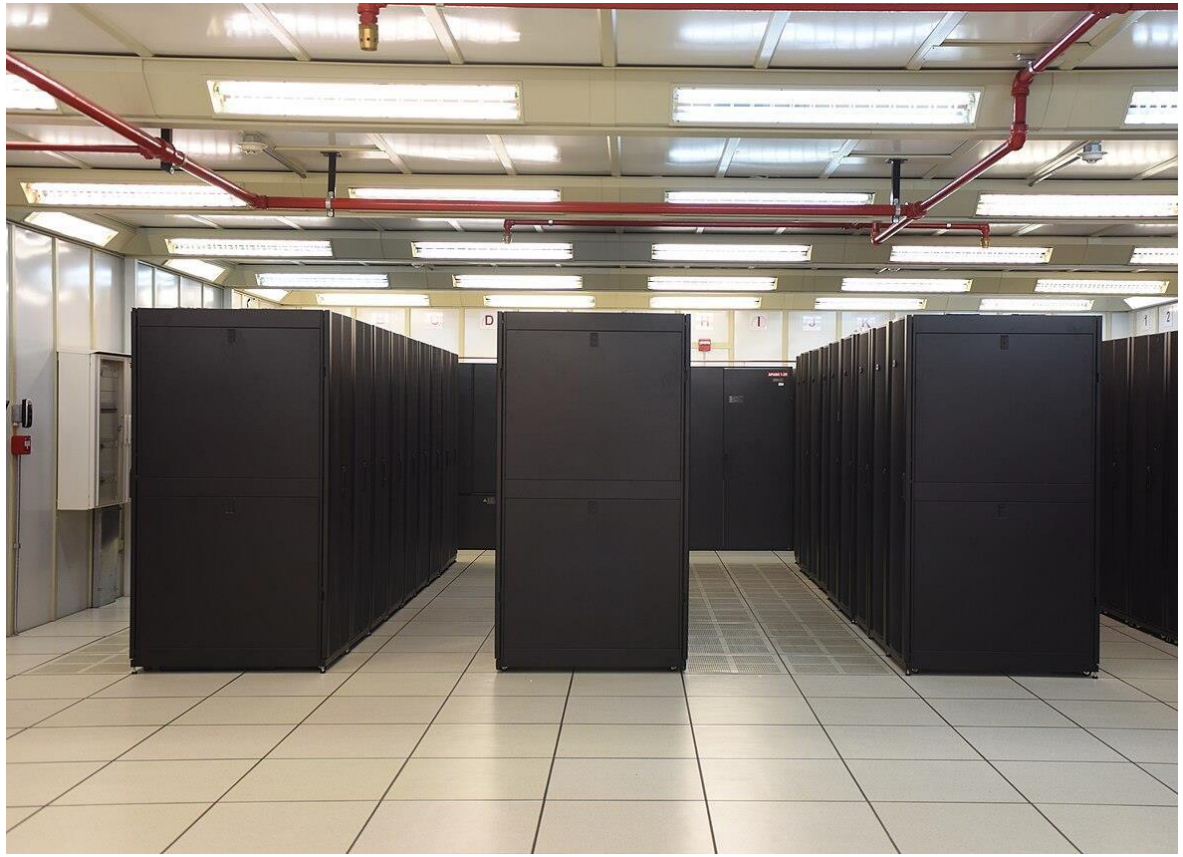


Image Credit: Data Center, Wikipedia Article (https://en.wikipedia.org/wiki/Data_center)

Model compression is an active area of research

Lots of interest in models that can get similar performance with fewer parameters

Big difference in speed and training time between a 70B model and a 1B model!

Business-wise, this saves major $$ in compute

Things can work on edge devices

- Robots
- Cell phones
- Car consoles

# Distillation – Where we left off

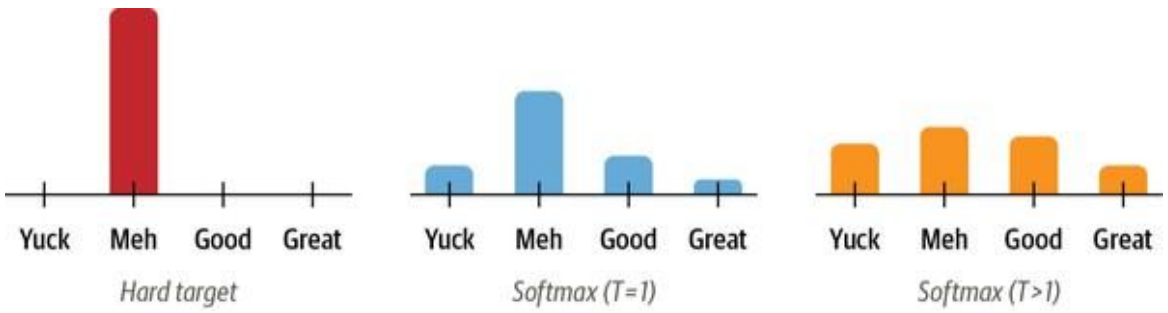



Image credit: *Natural Language Processing with Transformers*
Link to O'Reilly (free with Northeastern!)

**Soft probabilities**: Smoothed softmax probabilities, evenly distributed

**Distillation loss:** Divergence from teacher

$$L_{KD} = T^2 D_{KL}$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

**Main hyperparameters**

- **Temperature**: Controls extent of smoothing
- **Alpha**: Strength of distillation loss

# Distilling BERT on GLUE

Image Credit: BERT - Muppet Wiki (https://muppet.fandom.com/wiki/Bert)

GLUE – General Language Understanding Evaluation

Two tasks selected:

- MNLI – Multi-Genre Natural Language Inference

Do two sentences logically follow each other (0: entailment, 1: neutral, 2: contradiction)

- QQP – Quora Question Pairs

Does question B match question A semantically?

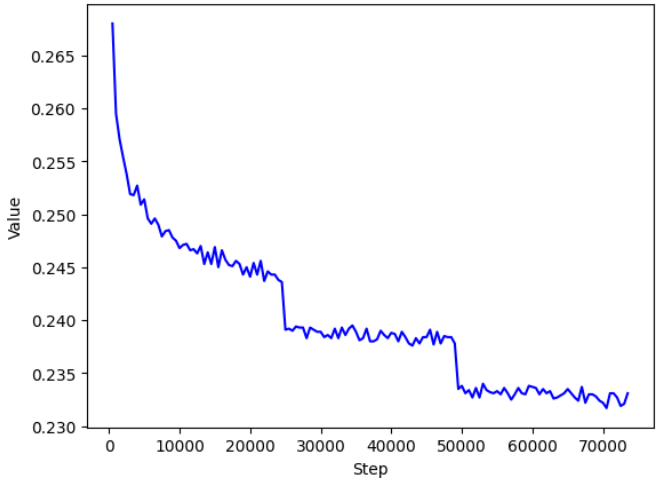
# Progress on Distilbert

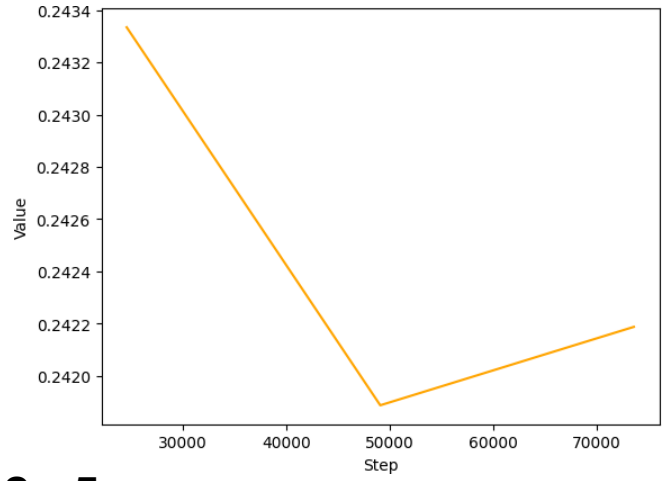
## Preliminary Results on MNLI

(F1 scores are macro)

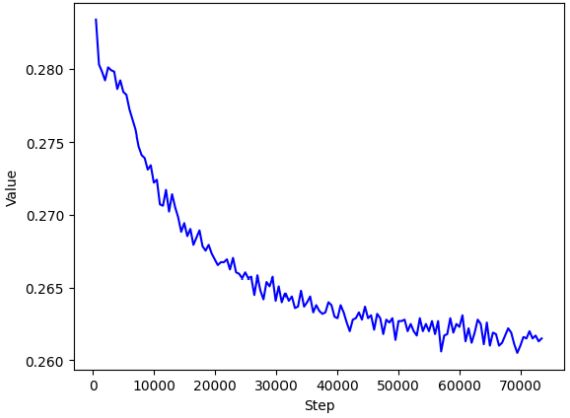
| Learning Rate | Accuracy | F1 |
|---|---|---|
| **2e-05** | **81.77%** | **81.68** |
| 2e-07 | 62.45% | 61.43 |
| 2e-09 | 31.99% | 17.93 |

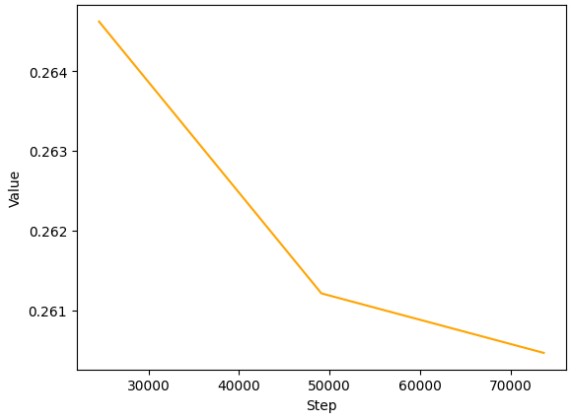
# MNLI Loss Curves



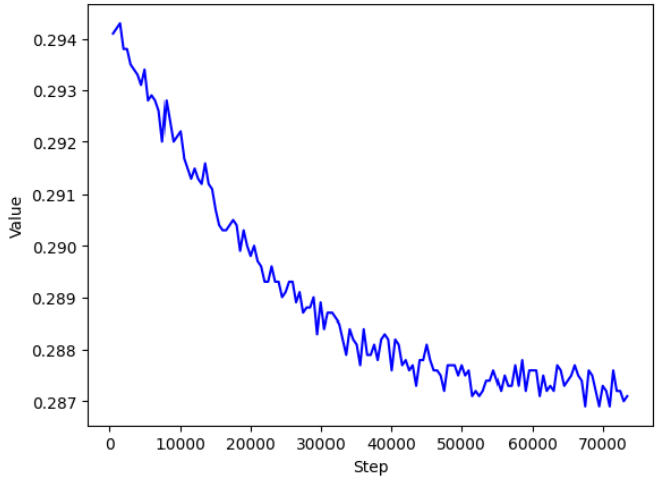



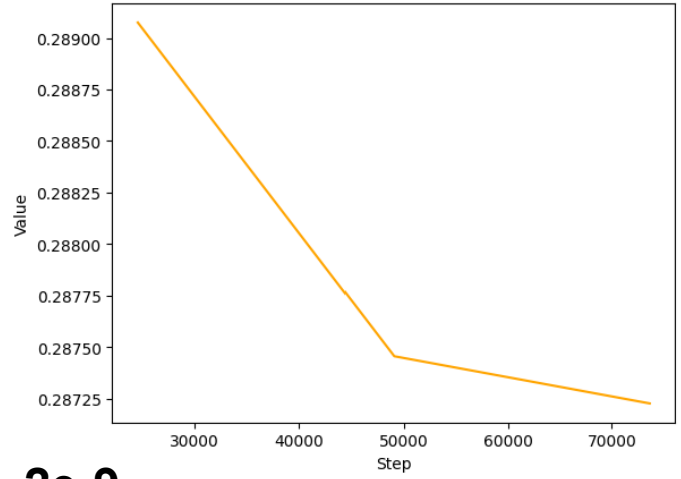

**2e-5**




**2e-7**



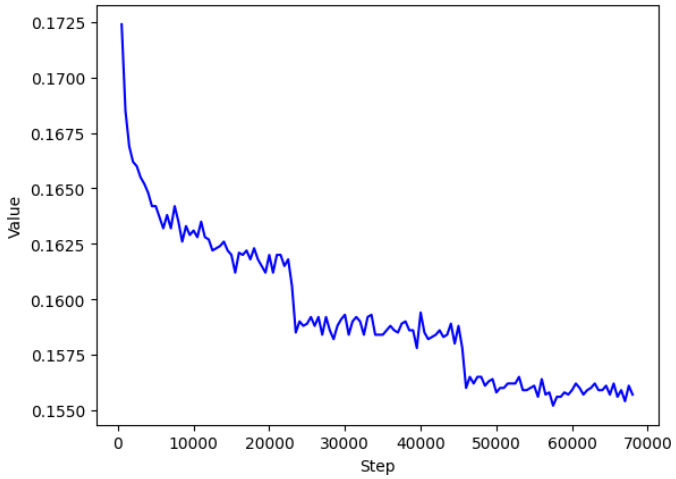

**2e-9**

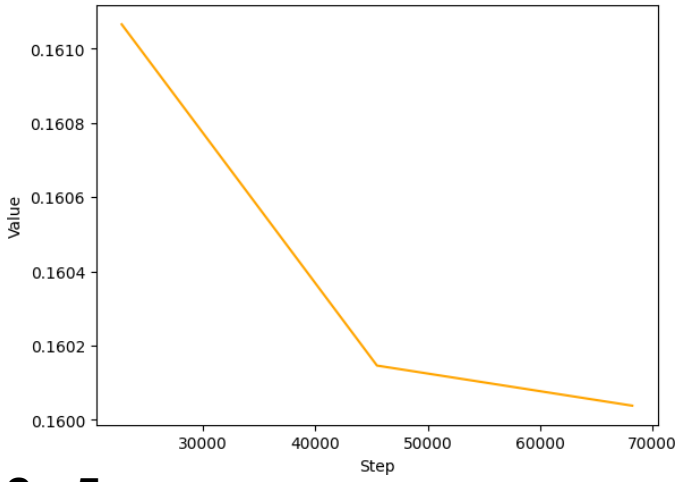
# Distilbert (cont'd)

Preliminary Results on QQP

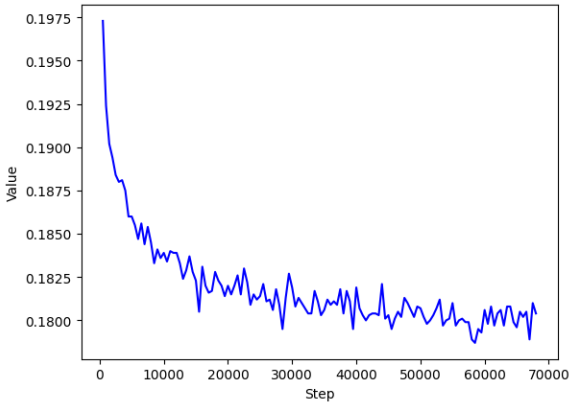
| Learning Rate | Accuracy | F1 |
|---|---|---|
| **2e-05** | **90%** | **89.46** |
| 2e-07 | 61.70% | 61.22 |
| 2e-09 | 63.18% | 38.72 |

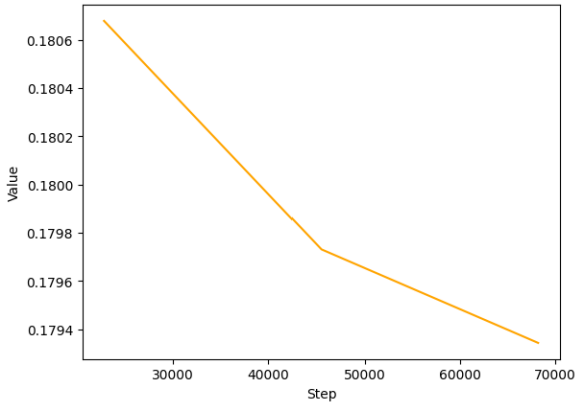
# QQP Loss Curves



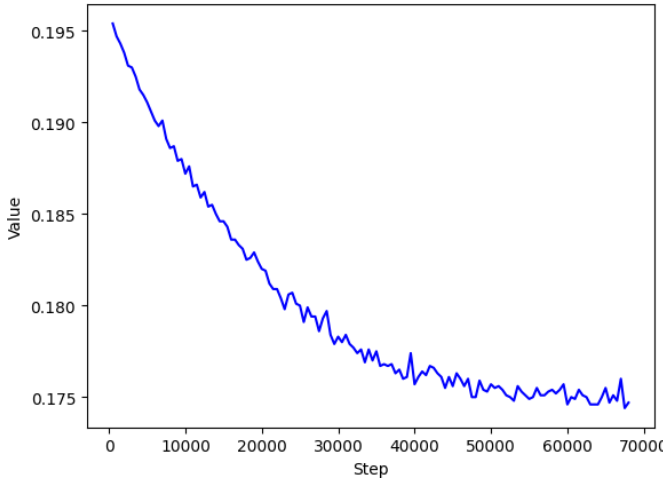



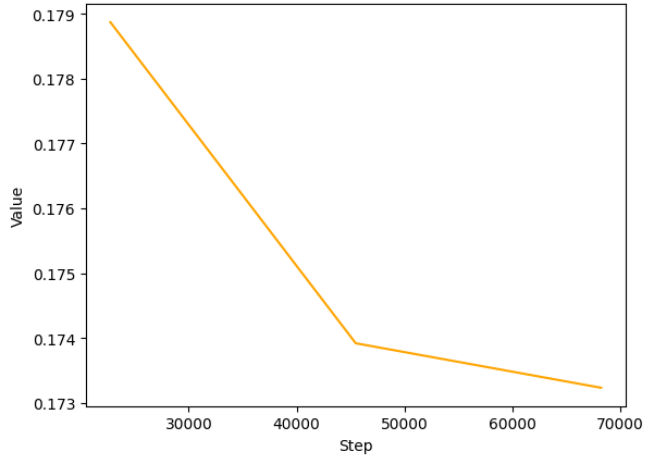

**2e-5**




**2e-7**




**2e-9**

# Introducing EuroLlama (patent pending)


We fine-tune the SOTA Llama3.2 on translation

- Decoder only translation (active research)
- Using distillation loss
- 3B checkpoint -> 1B
- Trained on a subset of Europarl (Koehn 2005)

Image credit: يعونصم شوه يفوعم امال‌ى Llama 3 3

# Progress on eurollama

**Chosen prompt and format:**

<|begin_of_text|><|start_header_id|>system<|end_header_id|>\n\nCutting Knowledge Date: December 2023\nToday Date: 31 Mar 2025\n\n You are a professional translator. Translate the provided text from English to French, remaining true to the source text. Do not add any additional commentary or conversational elements to your response.<|eot_id|><|start_header_id|>user<|end_header_id|>\n\nHello<|eot_id|><|start_header_id|>assistant<|end_header_id|>\n\nBonjour **<|end_of_translation|>**<|eot_id|>

# MT Results

Good BLEU: 30+

Good COMET: **0.80+**

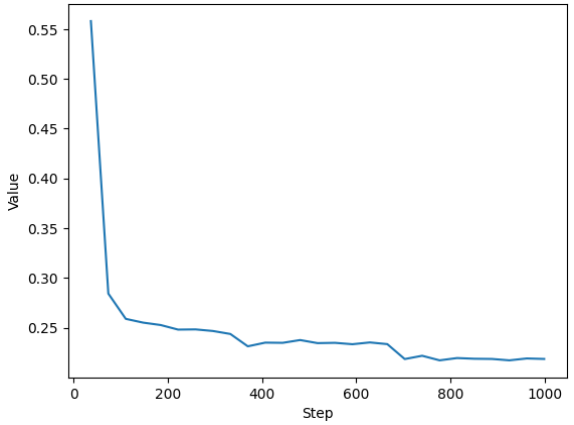
COMET uses multilingual embeddings

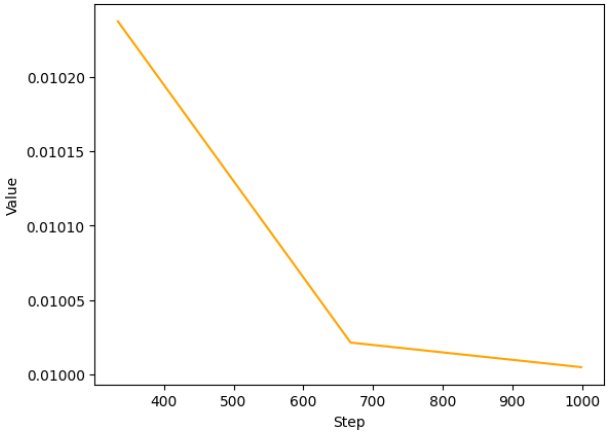
| Learning Rate | SacreBLEU | COMET |
|---|---|---|
| Base model | 18.89 | 0.73 |
| 1e-03 | 25.76 | 0.80 |
| **1e-04** | **26.57** | **0.82** |
| 1e-05 | 23.73 | 0.79 |

Legend:
Prompt
Source text
Target text

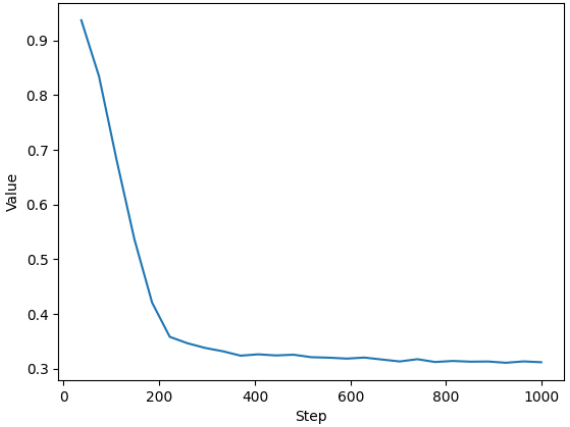
# Eurollama cont’d

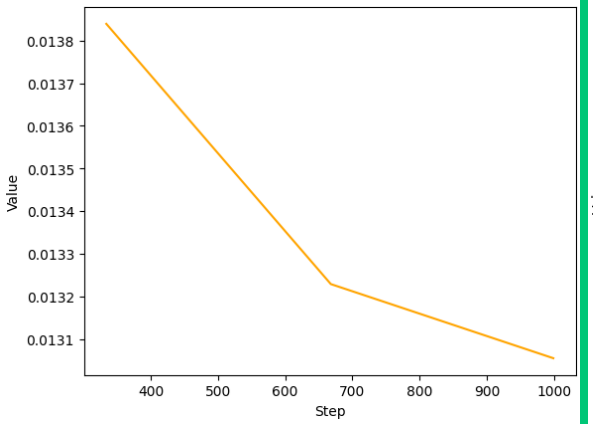
## LR: 1e-03



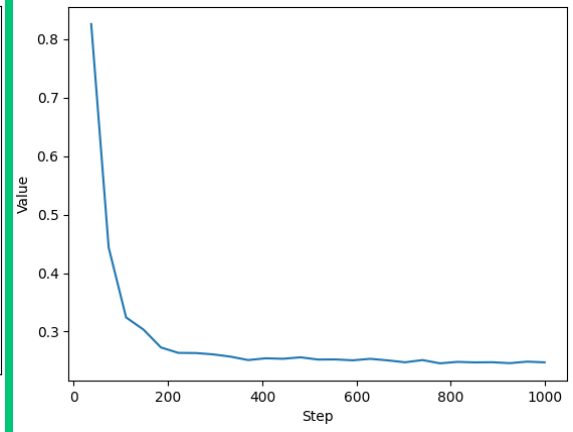



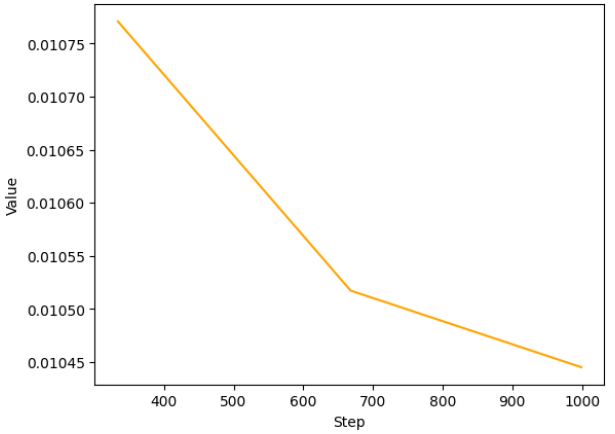

## LR: 1e-05




## LR: 1e-04



# Translation Limits and Snafus

**Translation**

- Quantization caused severe performance degradation
- Needed to use full precision on research Discovery cluster for translation training (P100 had insufficient VRAM)
- Time-intensive training, 8 hours only gets through 30% of dataset
- Hallucinated tokens, needed to create a new token!! <|end_of_translation|> (someone call Meta and let them know)

# Findings

- Distillation is a powerful technique!
- But there are additional hyperparameters to consider
- Hugging Face helps us implement this with modular Trainer class

**DistilBERT Findings**

- Higher alpha = higher evaluation accuracy
- 0.5 alpha sweet spot, balanced KL/CE loss

**Eurollama Findings**

- Distillation + encoder-only powerful for translation!
- KL Divergence good for causal loss