



Foundations of LLM Mastery: Fine-tuning with multi GPUs

25 February 2025
ONLINE



Evaluation

Or: how to take a Llama's measurements

Speaker: Simeon Harrison
Trainer at EuroCC Austria

Measuring up

Huggingface evaluate library: <https://huggingface.co/docs/evaluate/>

Different types of evaluation, depending on

- Goals
- Datasets
- Models

Huggingface currently offers/supports

- Metrics
- Comparisons
- Measurements



Generic

Metrics that can be applied to many different tasks and datasets.

- accuracy, precision, recall
- F1 score

Task specific

Tasks such as Machine Translation or Summarisation have specific metrics.

- BLEU
- ROGUE

Dataset specific

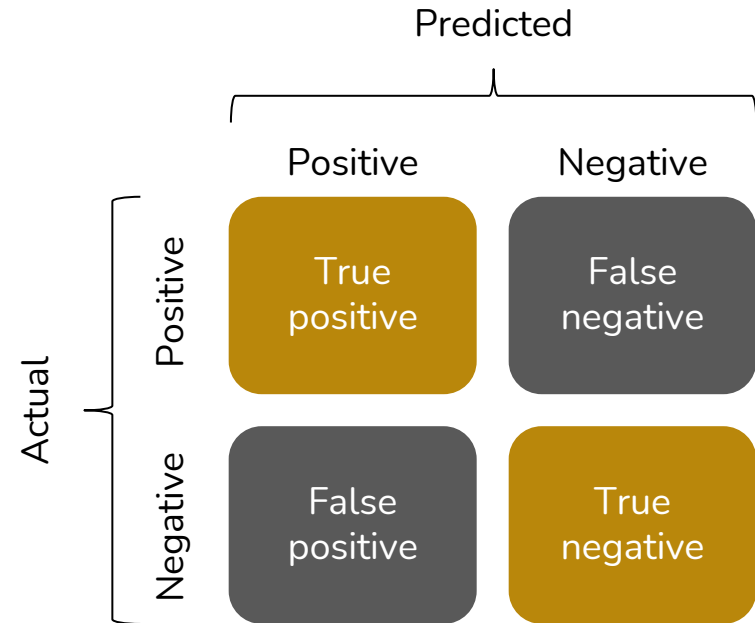
Datasets, which are used as benchmarks use specific metrics.

- GLUE
- SQuAD

| Metrics

Confusion Matrix

Generic



Generic

Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision

$$precision = \frac{TP}{TP + FP}$$

Recall

$$recall = \frac{TP}{TP + FN}$$

F1 score

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Task specific

BLEU

Algorithm for evaluating the quality of **machine-translated text**.

Quality is considered to be the correspondence between a machine's output and that of a human:

“The closer a machine translation is to a professional human translation, the better it is”

Source: Huggingface website

Task specific

ROUGE

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics used for evaluating **summarizations performed by LLMs**

The metrics compare a generated summary or translation against a reference human-produced summary.

GLUE

The **GLUE (General Language Understanding Evaluation)** benchmark is a collection of tasks designed to evaluate and benchmark natural language understanding models across a variety of language tasks.

Rather than a single metric, GLUE includes multiple tasks, each with its own evaluation metrics. These tasks help assess different aspects of language understanding, such as sentiment analysis, natural language inference, and linguistic acceptability.

Dataset specific

Dataset specific

SQuAD

The SQuAD (Stanford Question Answering Dataset) metric is specifically designed to evaluate Question Answering (QA) systems, particularly for tasks where a model answers questions based on a given context or passage.

The primary metrics used in the SQuAD dataset are:

- Exact Match
- F1 Score

Useful to compare the performance of several models on a given test dataset.

Comparisons are not used a lot yet.

Example: McNemar Test

Paired, nonparametric statistical hypothesis test. Compares predictions of two models and measures their divergence.

A vertical orange bar is positioned to the left of the section header.

Comparisons

Used to gain more insight on datasets and model predictions.

Datasets: average word length and distribution

Model predictions: average perplexity

A short, thick vertical gold bar.

Measurements

THANK YOU



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia