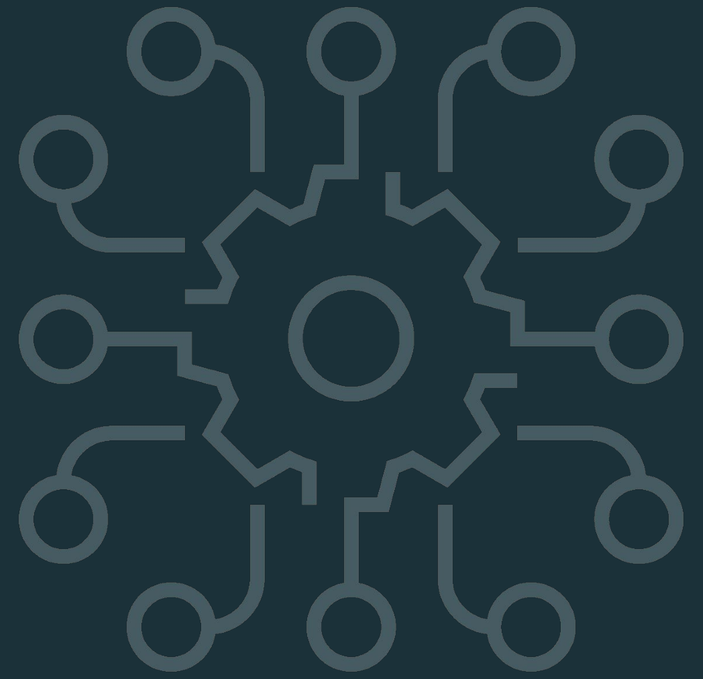
databricks
Academy

Evaluating LLMs



Databricks Academy
2023

Learning Objectives

By the end of this module, you should be able to:

- Compare and contrast the evaluation of traditional ML models and LLMs
- Understand how LLMs are generally evaluated, using a variety of metrics.

Evaluating LLMs vs. Traditional ML Models

Data and Resource Requirements

Traditional ML Models:

- Can be trained on less resource-intensive hardware.

LLMs:

- Requires massive amounts of data and substantial computational resources (GPUs, TPUs).

Evaluation Metrics

Traditional ML Models:

- Evaluated by metrics (F1, accuracy, etc.) focused on specific tasks like classification and regression.

LLMs:

- Evaluated using language specific metrics (BLEU, ROUGE, perplexity).
- Metrics are used to measure the quality of generated content.

Interpretability

Traditional ML Models:

- Often provide interpretable coefficients and feature importance scores.

LLMs:

- Especially large models seen as “black boxes” with limited interpretability.

Evaluating LLMs:

Overview of Evaluating LLMs

Training Loss/Validation Scores

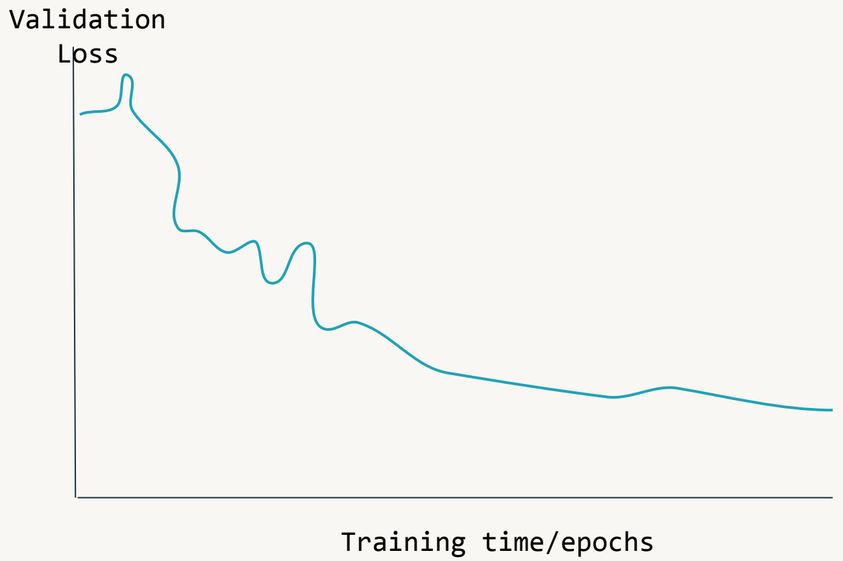
What we watch when we train

Like all deep learning models, we monitor the loss as we train LLMs.

But for a good LLM, what does the loss tell us?

Nothing really. Nor do the other typical metrics

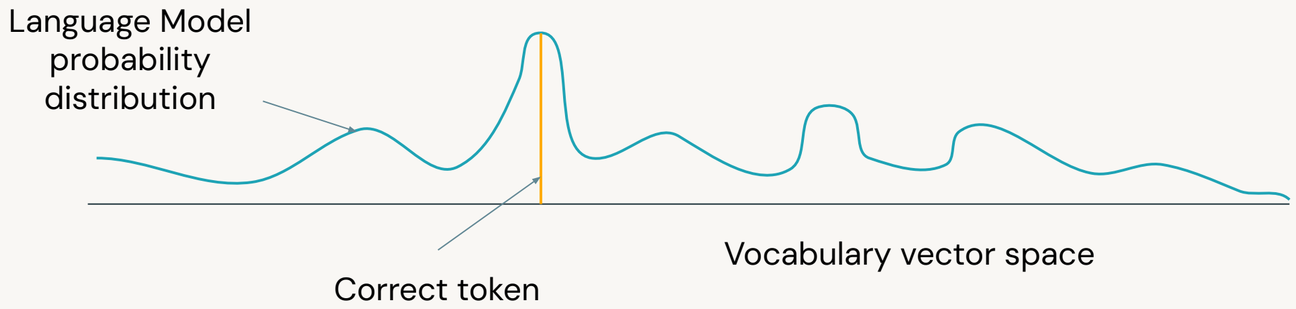
Accuracy, F1, precision, recall, etc.



Perplexity

Is the model surprised it got the answer right?

A good language model will have **high accuracy** and **low perplexity**



Accuracy = next word is right or wrong.

Perplexity = how confident was that choice.

More than perplexity

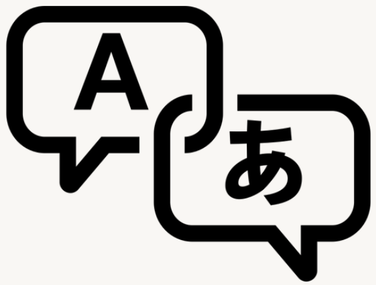
Task-specific metrics

Perplexity is better than just accuracy.

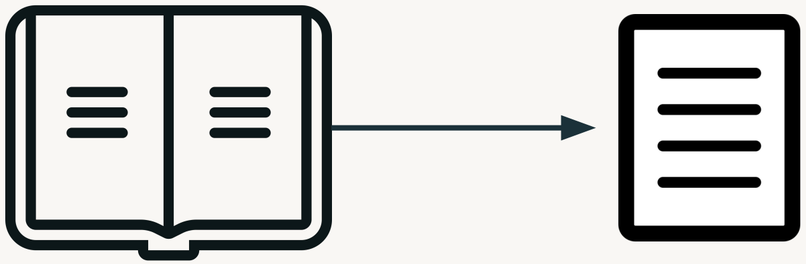
But it still lacks a measure of context and meaning.

Each NLP task will have different metrics to focus on. We will discuss two:

Translation – BLEU



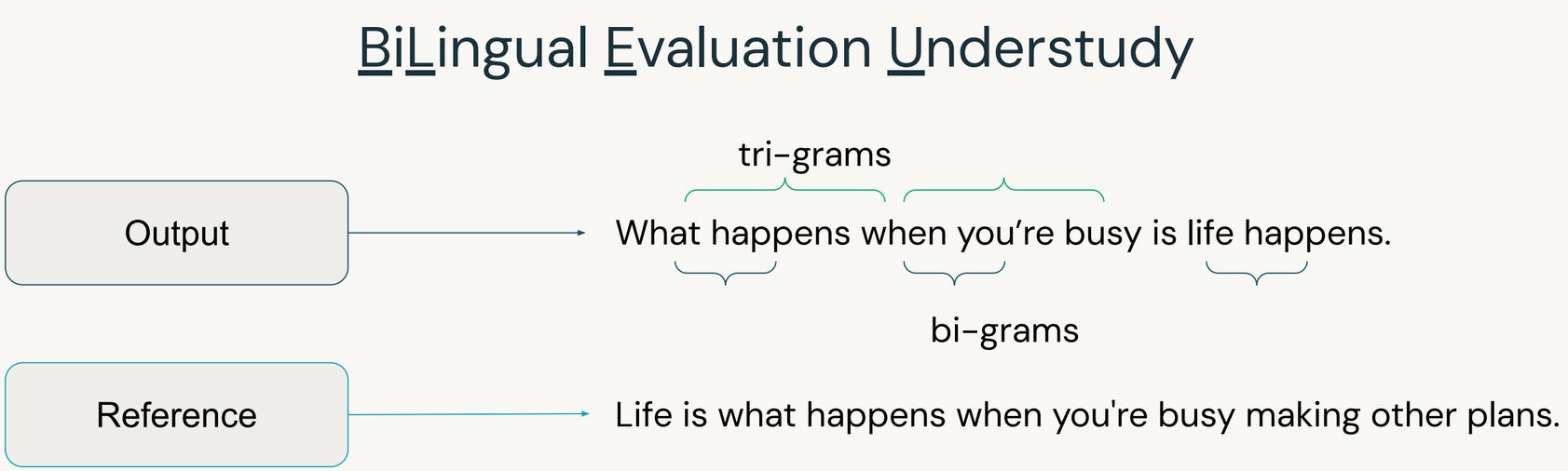
Summarization – ROUGE



Evaluating LLMs:

Task Specific Evaluations

BLEU for translation



BLEU uses reference sample of translated phrases to calculate n-gram matches: uni-gram, bi-gram, tri-gram, and quad-gram.

ROUGE for summarization

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference summaries}\}} \sum_{gram_n \in S} Count_{match}\left(gram_n\right)}{\sum_{S \in \{\text{Reference summaries}\}} \sum_{gram_n \in S} Count\left(gram_n\right)}$$

Total matching N-grams

Total N-grams

N-gram recall

ROUGE score for N-grams, e.g., ROUGE-1 for words

Sum over reference summaries (test data)

Sum over N-grams in summary *S*

ROUGE-1	Words (tokens)
ROUGE-2	Bigrams
ROUGE-L	Longest common subsequence
ROUGE-Lsum	Summary-level ROUGE-L

Reference: https://aclanthology.org/W04-1013.pdf

Benchmarks on datasets: SQuAD

Stanford Question Answering Dataset - reading comprehension

- Questions about Wikipedia articles
- Answers may be text segments from the articles, or missing

Given a Wikipedia article

Steam engines are external combustion engines, where the working fluid is separate from the combustion products. Non-combustion heat sources such as ***solar power***, nuclear power or geothermal energy may be used. The ideal thermodynamic cycle used to analyze this process is called the Rankine cycle. In the cycle, ...

Given a question

Along with geothermal and nuclear, what is a notable non-combustion heat source?

Select text from the article to answer (or declare no answer)

"solar power"

References: Rajpurkar et al., 2016 and https://rajpurkar.github.io/SQuAD-explorer/

Evaluation metrics at the cutting edge

ChatGPT and InstructGPT (predecessor) used similar techniques

1. Target application
    a. NLP tasks: Q&A, reading comprehension, and summarization
    b. Queries chosen to match the API distribution
    c. Metric: human preference ratings
2. Alignment
    a. “Helpful” → Follow instructions, and infer user intent. Main metric: human preference ratings
    b. “Honest” → Metrics: human grading on “hallucinations” and TruthfulQA benchmark dataset
    c. “Harmless” → Metrics: human and automated grading for toxicity (RealToxicityPrompts); automated grading for bias (Winogender, CrowS-Pairs)
        i. Note: Human labelers were given very specific definitions of “harmful” (violent content, etc.)

Reference: https://arxiv.org/abs/2203.02155

Evaluating LLMs:

Evaluation Challenges

Challenges of Evaluating LLMs

Lack of Ground Truth

- Generated text may not always align with human judgment or domain-specific knowledge.
- Evaluating subjective tasks like text generation is challenging.

Evaluation Metrics

- Evaluation metrics like BLEU and ROUGE, measure fluency but not quality of generated content.
- Metrics for evaluating aspects like coherence, relevance, and factual accuracy are still missing.

Ethical and Bias Concerns

- LLMs can generate biased or harmful content
- Detecting and mitigating bias is challenging.

Offline Evaluation

Evaluate based on human or another LLM labeling

Step 1: Curate a benchmark dataset

Curate benchmark datasets to measure various aspects of language generation, such as fluency, coherence, and grammaticality.

Step 2: Use metrics for text similarity

Use metrics like BLEU, ROUGE, perplexity, and F1-score to assess text similarity, language fluency, and token-level accuracy.

Step 3: Evaluate results

Evaluate metrics:

- Human evaluation
- Use another LLM to auto-evaluate results

Online Evaluation

Evaluate based on user behavior statistics. Data sources:

A/B Testing

LLMs are integrated into real-world applications, and their performance is assessed through A/B testing.

Direct Feedback

Collect user's direct feedback for the generated text.

Example: User rating and comments

Indirect Feedback

Collect user's indirect feedback based on their behavior.

Example: Clicks and conversions

Demo

Evaluating LLMs

Outline

- Evaluation metrics
- ROUGE score
  - ROUGE calculation for summarization task
  - Interpreting ROUGE scores
  - Comparing various models(t5-small, t5-base, gpt-2)

Lab

Evaluating LLMs

Outline

- Data Preparation
- Translation with LLMs
  - Translation with T5-Small
  - Translation with Helsinki-NLP
- Computing BLEU Score
- Model Comparison and BLEU Score Interpretation

databricks
Academy

Module Summary and Next Steps

Databricks Academy
2023

Module Summary

Let’s review

- Evaluating a model is crucial for model efficacy testing.
- Generic evaluation tasks are good for all models.
- Each NLP task will have different metrics to focus on.
- Specific evaluation tasks related to the LLM focus are best for rigor.

Helpful Resources

Resources and tools for evaluating LLMs

- Evaluation and Alignment in LLMs
  - HONEST
  - LangChain Evaluate
  - OpenAI's post on InstructGPT and Alignment
  - Anthropic AI Alignment Papers