5 WAYS TO → EVALUATE LLM RESPONSES

Bhavishya Pandit

WHY EVALUATE IN THE FIRST PLACE?

Although LLM models are trained on preset datasets, still they need to be checked for their responses for a couple reasons.

You might have heard about LLMs being biased in their responses, due to such reasons it is necessary to keep track of the responses these models create. There are other reasons involved as well:

- Real World Validation
- Bias & Ethical Oversight
- Ensuring Optimal Quality
- Boosting User Experience

1 PERPLEXITY

Perplexity measures the uncertainty of a language model's predictions. Simply put, it quantifies how well the model predicted probability distribution aligns with the actual distribution of the words in the dataset.

Imagine we have a language model trained to predict the next word in famous novels. If this model has a low perplexity when tested on a new, unseen novel, it suggests that its predictions closely match the actual word distributions. A model with low perplexity is likely more reliable as it can accurately predict the next word or token in a sequence.

$$P = b^{-1/N} \sum_{i=1}^{N} log_b p(w_i)$$

A lower perplexity indicates a model's better performance in predicting the sequence.

BLEU

BLEU is used for assessing generated text in various NLP areas by determining the n-gram overlap between the produced text and reference texts. BLEU's score depends on the n-gram precision in the produced text relative to the reference.

Let’s imagine a model generated “The sun rises in the east.” while we expected a reference Text as: “Sunrise is always in the east.”
Let’s calculate the BLEU score:

- Determine n-gram precision:
- Generated unigrams: [“The”, “sun”, “rises”, “in”, “the”, “east”, “.”]
- Reference unigrams: [“Sunrise”, “is”, “always”, “in”, “the”, “east”, “.”]
- Common unigrams = 5 (“sun”, “in”, “the”, “east”, “.”)
- Precision = Common unigrams / Total unigrams in produced text = 5/7 = 0.714.
- Brevity Penalty = Min(1, Words in generated text / Words in reference) = Min(1, 7/7) = 1.
- Final BLEU Score = Brevity Penalty exp(log(precision)) ≈ 1 exp(log(0.714)) ≈ 0.714 or 71.4%.

BLEU score ranges from 0 to 1. A high BLEU score indicates better translation quality

Bhavishya Pandit

ROUGE

ROUGE (Recall-Oriented Understudy for Gissing Evaluation) is another matrix used in text summarization assessment. It prioritizes recall over precision, measuring how much of the reference content is in the produced text.

ROUGE ensures exactness by checking linguistic similarity and correctness.

Let’s calculate the ROUGE score of our last sentence:

- Unigrams overlap: 5 (“sun”, “in”, “the”, “east”, “.”)
- Bigrams overlap (Generated): ["The sun", "sun rises", "rises in", "in the", "the east", "east ."]
- Bigrams overlap (Reference): ["Sunrise is", "is always", "always in", "in the", "the east", "east ."]
- Common Bigrams: 2 (“in the”, “the east”)
- ROUGE-1 Recall = Common unigrams / Total unigrams in reference text = 5/7 = 0.714.
- ROUGE-2 Recall = Common bigrams / Total bigrams in reference text = 2/6 = 0.333.
- ROUGE Score = (ROUGE-1 Recall + ROUGE-2 Recall) / 2 = (0.714 + 0.333) / 2 ≈ 0.524 or 52.4%.

Symbolizing that there is moderate similarity between the texts.

Bhavishya Pandit

4 METEOR

METEOR takes a multifactored approach to evaluating translations by considering accuracy, synonymy, stemming, and word order.

METEOR combines exact matches, stemmed matches, and paraphrase matching, and Its overall score is the harmonic mean of these factors.

Let’s calculate METEOR score of our sentence:

- Identify exact matches: 5 (“sun”, “in”, “the”, “east”, “.”).
- Stemmed matches: 0 (assuming basic stemming).
- Determine alignment: 0.714.
- Compute METEOR:
- Assuming Weight = 0.85 and Penalty = 0.775:
- METEOR = (1 — (0.85 (1–0.714))) 0.775 = (1–0.0429) * 0.775 ≈ 0.742 or 74.2%.

High METEOR score indicates high translation quality.

Bhavishya Pandit

5 ZERO SHOT

Metrics like Perplexity or accuracy only test the model on specific datasets thats where zero-shot is used.

Zero-shot learning refers to the ability of a model to understand and perform tasks it has never seen during its training phase. In the context of large language models, zero-shot evaluation means assessing the model's capability to handle prompts or questions not explicitly represented in the training data.

This metric helps in evaluating the following traits of a model:

1. Safety
2. Intelligence & Capability
3. Reliability

FOLLOW FOR MORE AI/ML POSTS

Bhavishya Pandit