POST 27 - GEN AI

IN-DEPTH ANALYSIS: KEY PERFORMANCE INDICATORS FOR VALIDATING LARGE LANGUAGE MODELS

Generative AI Deep Dives,
Model Validation KPIs

GENERATIVE AI FOR ALL

DINESH LAL
(DIRECTOR, DATA SCIENCE)

DINESHLAL

PERPLEXITY

What it Measures?

Predictive performance of the model on a sample of text.

Explanation

Perplexity evaluates how well a language model predicts a sequence of words. It's the exponentiation of the average negative log-likelihood of a set of words.

What Does Its Value Imply?

Lower perplexity values indicate better predictive performance.

Formula

: $e^{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i)}$

Example Values

- Good Model: 10-20
- Poor Model: 100

DINESHLAL

BLEU SCORE (BILINGUAL EVALUATION UNDERSTUDY)

What it Measures?

Quality of machine-generated text compared to reference text.

Explanation

BLEU score compares generated text to one or more reference texts by evaluating n-gram precision and applying a brevity penalty to discourage short outputs.

What Does Its Value Imply?

Higher BLEU scores indicate better quality and closer alignment with reference texts.

Formula

- $\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$
- **BP** (Brevity Penalty): $\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$
- w_n: Weights, usually uniform.
- p_n: Precision for n-grams.

Example Values

- Good Model: 0.5-1.0
- Poor Model: 0.1-0.3
-

DINESHLAL

ROUGE SCORE (RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION)

What it Measures?

Overlap of n-grams between generated text and reference text.

Explanation

ROUGE score evaluates n-gram overlap between generated and reference texts, focusing on recall.

What Does Its Value Imply?

Higher ROUGE scores indicate better text quality and alignment with reference texts.

Formula

$$\frac{\sum_{\text{matches} \in \text{hypothesis}} \text{Count}(\text{match N-grams})}{\sum_{\text{N-grams} \in \text{reference}} \text{Count}(\text{N-grams})}$$

Example Values

- Good Model: 0.6-0.8
- Poor Model: 0.2-0.4

DINESHLAL

PRECISION

What it Measures?

Proportion of relevant instances among the retrieved instances.

Explanation

Precision evaluates how many of the retrieved instances are actually relevant.

What Does Its Value Imply?

Higher precision values indicate fewer false positives.

Formula

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Example Values

- Good Model: 0.8-1.0
- Poor Model: 0.5-0.7

DINESHLAL

RECALL

What it Measures?

Proportion of relevant instances that were retrieved.

Explanation

Recall evaluates how many relevant instances were retrieved by the model.

What Does Its Value Imply?

Higher recall values indicate fewer false negatives.

Formula

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Example Values

- Good Model: 0.8-1.0
- Poor Model: 0.5-0.7

DINESHLAL

F1 SCORE

What it Measures?

Harmonic mean of precision and recall.

Explanation

F1 Score provides a single measure of performance by balancing precision and recall.

What Does Its Value Imply?

Higher F1 scores indicate better overall performance.

Formula

$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example Values

- Good Model: 0.8-1.0
- Poor Model: 0.5-0.7

DINESHLAL

LATENCY

What it Measures?

Time taken by the model to generate a response.

Explanation

Latency measures the responsiveness of the model.

What Does Its Value Imply?

Lower latency values indicate faster responses.

Formula

Total Time Taken/
Number of Requests

Example Values

- Good Model: < 100ms
- Poor Model: > 500ms

DINESHLAL

THROUGHPUT

What it Measures?

Number of requests the model can handle in a given period.

Explanation

Throughput measures the efficiency of the model in handling multiple requests.

What Does Its Value Imply?

Higher throughput values indicate better efficiency.

Formula

Number of Requests/
Total Time Taken

Example Values

- Good Model: 1000+ requests/second
- Poor Model: < 100 requests/second

DINESHLAL

MEMORY USAGE

What it Measures?

Amount of memory required during inference.

Explanation

Memory usage indicates how much memory the model consumes during operation.

What Does Its Value Imply?

Lower memory usage is preferable for scalability and efficiency.

Formula

- Measured using memory profiling tools during inference.

Example Values

- Good Model: < 1GB

EQUALIZED ODDS

What it Measures?

Difference in true positive and false positive rates across different demographic groups.

Explanation

Equalized Odds ensures that the model performs equally well across different groups.

What Does Its Value Imply?

Lower values indicate fairer models with less bias.

Formula

$$|\text{TPR}_{\text{group1}} - \text{TPR}_{\text{group2}}| + |\text{FPR}_{\text{group1}} - \text{FPR}_{\text{group2}}|$$

Example Values

- Good Model: < 0.1
- Poor Model: > 0.3

DINESHLAL

DEMOGRAPHIC PARITY

What it Measures?

Checks if different groups receive positive outcomes at the same rate.

Explanation

Demographic Parity ensures equal treatment across demographic groups.

What Does Its Value Imply?

Lower values indicate less bias.

Formula

$$| P(\text{outcome} = 1 \mid \text{group1}) - P(\text{outcome} = 1 \mid \text{group2}) |$$

Example Values

- Good Model: < 0.1
- Poor Model: > 0.3
-

CALIBRATION

What it Measures?

Alignment of predicted probabilities with true probabilities.

Explanation

Calibration measures if predicted probabilities are accurate.

What Does Its Value Imply?

Lower ECE values indicate better calibration.

Formula

$$| P(\text{outcome} = 1 \mid \text{group1}) - P(\text{outcome} = 1 \mid \text{group2}) |$$

Example Values

- Good Model: < 0.05
- Poor Model: > 0.1

Thank You

SPECIAL THANKS TO CHATGPT, OPEN AI, COPILOT, GEMINI FOR THE SUPPORT ON CONTENT

DINESHLAL