LLM-as-a-Judge Evaluation Metrics

This document provides a detailed overview of evaluation metrics for assessing **LLM-as-a-Judge** systems in reading comprehension tasks. The criteria include **Agreement, Rank Correlation, Cohen's Kappa, Bias Analysis, and Robustness**.

1. Key Evaluation Metrics

1.1 Agreement

Objective: Measures the proportion of exact matches between the LLM's evaluation and the human evaluation.

Formula:

$$\text{Agreement} = \frac{1}{N} \sum_{i=1}^{N} I(\text{LLM}_i = \text{Human}_i)$$

Where:

- N is the number of answers.
- $I(\text{LLM}_i = \text{Human}_i)$ is 1 if the LLM's score matches the human's score for the i-th answer, otherwise 0.

1.2 Spearman's Rank Correlation

Objective: Measures the correlation between LLM and human rankings.

Formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

Where:

- d_i is the difference between the LLM and human ranking of answer i.
- N is the number of answers.

A higher ρ means greater alignment between LLM and human rankings.

1.3 Cohen's Kappa

Objective: Measures inter-rater reliability between LLM and human evaluators.

Formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where:

- P_o is the observed agreement between LLM and human scores.
- P_e is the expected agreement by chance.

Higher κ values indicate stronger agreement.

---

2. Bias Metrics

2.1 Position Bias

Objective: Checks if the position of a question in a list influences the LLM's score.

Formula:

$$\text{Position Bias Correlation} = \text{Correlation}(\text{Position}, \text{LLM Generated Score})$$

Where:

- **Position** is the question's index in a sequence.
- **LLM Generated Score** is the evaluation given by the LLM.

A strong correlation suggests that position influences scores.

2.2 Length Bias

Objective: Checks if longer answers receive higher scores.

Formula:

$$\text{Length Bias Correlation} = \text{Correlation}(\text{Length}, \text{LLM Generated Score})$$

Where:

- **Length** is the number of words in the answer.
- **LLM Generated Score** is the evaluation score assigned by the LLM.

A positive correlation suggests preference for longer responses.

---

3. Robustness (Adversarial Testing)

Objective: Measures whether the LLM's evaluation is stable against minor modifications.

Formula:

$$\text{Robustness} = 1 - \frac{\text{Variance of LLM Scores after Perturbations}}{\text{Original LLM Score Variance}}$$

Where:

- **Original Variance** is the variance in LLM scores before perturbation.
- **Perturbed Variance** is the variance after small changes to the input.

A higher **robustness score** suggests that evaluations remain stable under slight input modifications.

---

4. Summary of Metrics

Metric	Formula	Description
Agreement	$\text{Agreement} = \frac{1}{N}\sum_{i=1}^{N} I(\text{LLM}_i = \text{Human}_i)$	Measures how often LLM and human evaluators give identical scores.
Spearman's Rank Correlation	$\rho = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$	Evaluates ranking consistency between LLM and human evaluations.
Cohen's Kappa	$\kappa = \frac{P_o - P_e}{1 - P_e}$	Measures inter-rater reliability, adjusting for chance.
Position Bias	$\text{Correlation}(\text{Position}, \text{LLM Generated Score})$	Assesses whether the placement of a question affects scoring.
Length Bias	$\text{Correlation}(\text{Length}, \text{LLM Generated Score})$	Checks if longer responses get higher ratings.
Robustness	$1 - \frac{\text{Variance of Perturbed Scores}}{\text{Original Score Variance}}$	Tests the stability

Metric	Formula	Description
		of LLM evaluations under slight modifications.

---

5. Conclusion

This document provides a structured approach to evaluating **LLM-as-a-Judge** systems using fundamental and bias-related metrics. These evaluations ensure **fairness, reliability, and consistency** in automated judgment tasks.

This markdown file serves as a reference guide for implementing **LLM evaluation pipelines**. If you need additional details or modifications, let me know!