**Negative Rejection**

The **Negative Rejection Metric** is a custom evaluation metric designed to quantify how well a generated response captures the meaning of the ground truth (reference) response. It operates by measuring the similarity between the semantic embeddings of the generated and true responses and determining how often the generated response fails to capture the essential meaning of the true response.

**How the Metric Works**

The metric compares each generated response with its corresponding true response using a similarity threshold. If the similarity score is below this threshold, the generated response is considered to have **"rejected"** the true meaning, i.e., it failed to capture the essence of the true response.

The metric computes the **Negative Rejection Rate (NRR)** as the proportion of responses where the generated text did not meet the similarity criteria compared to the total number of responses evaluated.

**Steps Involved:**

1. **Tokenization and Encoding:**
   * Both the ground truth and generated responses are tokenized and encoded using a pre-trained language model such as **Sentence Transformers** or any transformer-based model.
2. **Embedding Generation:**
   * The responses are passed through a transformer model to generate embeddings, which represent the semantic meaning of the responses.
3. **Similarity Calculation:**
   * **Cosine Similarity** is used to measure the similarity between the generated response and the ground truth response.
4. **Threshold Comparison:**
   * The cosine similarity score is compared to a pre-defined similarity threshold (e.g., 0.8).
   * If the maximum cosine similarity between the generated and true responses is below the threshold, the response is classified as **"rejected"**.
5. **Negative Rejection Rate Calculation:**
   * The total number of "rejected" responses is divided by the total number of evaluated responses to compute the **Negative Rejection Rate (NRR)**.

NNR = Total Capture / (Total Capture + Total Not Captured