Step-by-Step Approach for Evaluation

Sample Selection:

- Selected 10 representative articles from the Kaggle Medium Articles Dataset for analysis.
- The titles and tags of the chosen articles were inputted into both the original (Baseline) and the fine-tuned versions of the GPT2 and LLAMA2 models to generate text for comparison.

Step 1: Evaluation Metrics Identification:

Chat GPT-4 was tasked with analyzing the writing style of the actual content of the sample Medium articles to pinpoint the 5 most critical metrics for evaluation.

Step 2: Weight Assignment for Metrics:

Chat GPT-4 was then asked to assign importance weights to these metrics with a maximum possible score of 10 per article to enable a consistent and stringent evaluation.

S No.	Metric	Description	Weight
1	Content Coherence and Relevance	Alignment with title and tags and thematic accuracy	2
2	Style and Tone Consistency	Appropriateness of writing style and tone for the subject and audience.	3
3	Engagement and Readability	Article's narrative flow, engaging elements, and ease of reading.	2
4	Originality and Creativity	Uniqueness of insights, perspectives, and approaches.	2
5	Technical Accuracy and Depth	Factual correctness and comprehensive analysis of specialized topics.	1

Step 3: Comparative Evaluation:

Used Chat GPT-4 to evaluate each article against the set metrics, documenting the performance of untuned (baseline) and fine-tuned models of both GPT2 and LLAMA2.

Step 4: Score Averaging and Differential Analysis:

Calculated the average scores for the baseline (untuned) models and the fine-tuned models with LORA for both GPT2 and LLAMA2, then analyzed the differences between these scores to determine the impact of fine-tuning.