

The evaluation strategy for the fine-tuned LLMs focuses on assessing both training and validation performance to ensure the model's ability to generalize across unseen data. The training and validation losses were analyzed at different training steps, with results showing a consistent decrease in training loss (e.g., from 1.3387 at Step 5 to 0.4920 at Step 25), indicating effective learning. The validation loss also displayed a generally downward trend, although minor fluctuations were observed, such as an increase from 0.7255 to 0.6472 and stabilization around 0.5423. These observations were complemented by the evaluation results, where the model achieved an evaluation loss of 0.5423, an evaluation runtime of 4.9069 seconds, and an evaluation throughput of 16.304 samples per second. This strategy ensures that both training and unseen data performance are closely monitored to validate convergence and model robustness.

The evaluation methodology was selected to strike a balance between training efficiency and inference performance. A detailed step-wise evaluation allowed us to measure losses at each critical training milestone, ensuring timely checkpoints and convergence tracking. The evaluation sampling throughput (16.304 samples/second) and steps per second metrics (2.038) highlight the efficiency of the model's inference capabilities. Additionally, more robust evaluations, including domain-specific datasets and real-world text samples, would further confirm the model's generalizability and adaptability across diverse scenarios.