To evaluate the performance of the fine-tuned Icendanic\_Meta-Llama-3.1-8B model, I employed a comprehensive evaluation strategy involving multiple metrics. The primary evaluation metric used was evaluation loss, which provides a direct measure of the model's ability to predict the target output compared to the ground truth. A lower loss indicates that the model's predictions are closer to the true values. Additionally, training time was also recorded to assess the model's efficiency, and peak memory usage was monitored to ensure the model operates within acceptable memory limits during the training and evaluation process. The evaluation loss for the model was recorded at 0.8486, indicating room for improvement in terms of minimizing prediction errors. The evaluation accuracy was not reported, as BLEU score, a common metric for translation tasks, was found to be 0, indicating that the translation performance did not meet expectations during this evaluation phase.

The rationale behind selecting evaluation loss as the primary metric stems from its ability to provide direct insight into how well the model performs across the task's training data, while BLEU score was expected to provide a more nuanced view of the translation quality, especially for machine translation tasks. However, the lack of an improved BLEU score suggests the need for further model optimization or adjustments to hyperparameters such as the learning rate or fine-tuning strategies. The training and evaluation were conducted over a total of 1200 seconds, or 20 minutes, with peak memory usage reaching 15.75% of the maximum available system memory, ensuring efficient model operation. These metrics combined with further refinement strategies are essential for optimizing the model's performance in real-world translation tasks.



