Documentation: Choice of Unsloth Configuration and Training Strategy

For the fine-tuning of LLaMA 3.1-8B-Instruct on the 'Bengali Empathetic Conversations dataset', I selected unsloth as the primary fine-tuning strategy. The choice was guided by a need to balance model performance, memory efficiency, and feasibility on free GPU resources, such as those provided by Kaggle's T4 instances. To facilitate flexible experimentation, I implemented the 'Strategy Pattern', allowing us to dynamically switch between Unsloth and a standard LoRA-based PEFT approach. This pattern decouples the fine-tuning logic from the rest of the training pipeline, enabling reproducibility and modularity. With this design, swapping strategies requires only a single configuration change, without modifying the core model or training code.

The Unsloth configuration was optimized to maximize model capacity while keeping within GPU memory limits. LoRA adapters were applied to all critical attention projections the query, key, value, and output layers to capture the aspects of contextual and semantic understanding. The rank parameter (r = 16) was chosen to provide sufficient modeling capacity, while the scaling factor (lora_alpha = 32) ensures stable gradient updates. A small dropout (lora_dropout = 0.05) was used to reduce overfitting, and bias parameters were frozen to minimize unnecessary trainable parameters. Most importantly, Unsloth's gradient checkpointing was employed to allow the model to process full-sequence tokenization of up to 4096 tokens, preserving the complete context of empathetic conversations, which is crucial for generating meaningful responses.

The training strategy was designed to efficiently use limited GPU resources while maintaining high-quality fine-tuning. The dataset was carefully processed to include 6,000 curated Bengali conversation samples, with full-sequence tokenization and proper masking of prompt tokens to support teacher-forced training. To cope with memory constraints, I set a per-device batch size of 1 and accumulated gradients over four steps to emulate a larger batch. Mixed-precision training (fp16) further optimized GPU memory usage. The model was trained for 2 epochs as an initial baseline, with checkpoints and logs saved regularly to ensure reproducibility.

Evaluation was incorporated directly into the pipeline, using standard automatic metrics such as Perplexity, BLEU, and ROUGE, as well as human evaluation prompts to assess the quality and empathy of generated responses. A memory-efficient data

collator was implemented to dynamically pad sequences and mask padding tokens during loss computation, ensuring that training was both effective and stable.

Evaluation Metrics Analysis:

The model achieved a perplexity of 1.97, which is quite low, indicating that the fine-tuned LLaMA 3.1-8B-Instruct is effectively predicting the next token in the Bengali empathetic conversations. A lower perplexity generally reflects better language modeling performance and shows that the model has learned the overall structure and style of the dataset.

However, the BLEU score is very low (0.0085), and all ROUGE scores are zero. This suggests that while the model can generate fluent Bengali text, it does not match the reference answers exactly. This is expected in tasks like empathetic conversation generation because there are often many valid ways to respond to a given prompt — strict n-gram matching metrics like BLEU and ROUGE may not fully capture the quality or empathy of the responses.

In other words, these automatic metrics indicate limited overlap with reference texts, but they do not necessarily reflect the actual human-perceived quality of the generated responses.

Challenges Faced:

During the training process, several challenges were encountered that impacted the overall workflow and final model performance. Initially, attempting LoRA-based fine-tuning led to out-of-memory (OOM) errors on Kaggle's GPU, which necessitated switching to the Unsloth approach for memory-efficient parameter updates. Training on the full dataset proved time-intensive, exceeding Kaggle's 12-hour session limit, so only a smaller subset of the data could be used. Memory constraints also limited training to a very small fraction (around 0.17%) of LLaMA's parameters, which contributed to less optimal automatic evaluation scores. Additionally, the GPU occasionally crashed or paused spontaneously, requiring retraining and further extending the total training time. Despite these limitations, the model was successfully fine-tuned to generate coherent Bengali empathetic responses, demonstrating the feasibility of parameter-efficient training under constrained resources.