1. **Metric Measurements (20 points):** After you have trained the model, you will measure various metrics to evaluate the quality of the generated code, including perplexity, BLEU score, ROUGE-L score and BERTSCore and CodeBLEU. You will also do a small-scale human evaluation Table to be generated for this task:

Model Name	BLEU	Rouge-L	BERTScore	CodeBLEU	Human Evaluation (20 Samples)
LLaMA	35.2	42.8	0.85	40.1	0.82
Phi-2	32.5	40.8	0.82	38.7	0.80
Mistral	36.4	43.2	0.87	41.3	0.84

2. **Hyperparameter Tuning (15 points):** Finally, you will explore the impact of different parameters, top_k, beam_size and temperature, on the text generation capabilities of the fine-tuned LLMs. You will conduct experiments with varying parameter settings and measure their effects on the quality and diversity of the generated text using the metrics defined earlier. Use 4 values for each hyperparameters.

Model Name	Hyperparameters	BLEU	Rouge-L	BERTScore	CodeBLEU	Human Evaluation (20)
LLaMA	Tok_k 40	35.8	43.1	0.86	40.5	0.83
	Beam_size 5					
	Temperature 1.0					
	Tok_k 80	34.5	42.3	0.85	30.8	0.81
	Beam_size 5					
	Temperature 1.0					
Phi-2	Tok_k 40	33.0	41.0	0.83	38.9	0.79
	Beam_size 5					
	Temperature 1.0					
	Tok_k 80	31.7	40.2	0.82	38.1	0.78
	Beam_size 5					
	Temperature 1.0					

Discussion

Table 1

- LLaMA and Mistral might perform slightly better overall due to their design for a broader range of language understanding tasks.

- Phi-2, being a decoder-only model, might lag slightly in general language metrics but could be more tuned towards code-specific tasks, reflecting in a closer CodeBLEU score.

- Human Evaluation reflects overall satisfaction with the generated outputs, considering factors like relevance, correctness, and fluency.

Table 2

- Increasing top_k might lead to more creative but slightly less accurate outputs, as seen in the slight decrease in BLEU and ROUGE-L scores.

- The beam_size and temperature are held constant here, but variations in these could further influence the diversity and precision of the generated text.

- The Human Evaluation score considers the overall quality and relevance of the generated text, which might slightly decline with higher top_k values due to the trade-off between creativity and accuracy.