Exploring Large Language Models (LLMs) for Text Generation with PyTorch and Hugging Face

Report

Model Name	BLEU	Rouge-L	BERTScore	CodeBLEU	Human Evaluation (20 Samples)
LLaMA	0.16929356226897385	0.23686005208844993	[0.8385592699050903 0.891447901725769 0.8635869026184082]	0.3446820336621457	0.96
Phi-2	0.11255576346039123	0.20615714923460185	[0.8342348337173462 0.874875009059906 0.8535738587379456]	0.295976400573006	0.65
Mistral	0.15507475287381808	0.2431139473896443	0.8381794095039368 0.8793331384658813 0.8576527237892151	0.3028048618318122	0.95

1. **Write a discussion (4-5 Lines) explaining the comparison between two models. Moreover, compare the metrics and discuss which metrics are more appropriate compared to human evaluation.**

   The comparison between LLaMA 2, Phi-2, and Mistral displays differences in performance across various metrics. LLaMA and Mistral outperform Phi-2 in BLEU, Rouge-L, BERTScore, and CodeBLEU, indicating the best performance in code generation understanding. However, during testing, it was observed that the Phi Model tends to give the correct answer, but after giving the correct answer, it keeps giving more “exercise sample” answers that tend to hallucinate. Mistral and Llama obtained similar evaluations on the human eval metric, reflecting a good understanding of code generation. While automatic evaluation metrics like BLEU, Rouge-L, BERTScore, and CodeBLEU provide quantitative understandings of model performance, they might not fully capture human judgment. Hence, depending exclusively on these metrics for code understanding may not always reflect the true quality of model outputs.

Model Name	Hyperparameters (top_k,beam_size,temperature)	BLEU	Rouge-L	BERTScore	CodeBLEU	Human Evaluation (20 Samples)
LLaMA	5,10,0.7	0.25912309811748196	0.27863458186442 89	0.858018517494 2017 0.890473544597 6257	0.3801029960921 649	0.8857

				0.873663306236 2671		
	25,5,0.8	0.2353444821154 9008	0.2590992675518 776	0.847023963928 2227 0.890600800514 2212 0.867829978466 0339	0.3613913368720 005	0.8785
	36,8,0.5	0.1614958074180 437	0.2520188395535 2666	[0.842839360237 1216 0.897470235824 585 0.868566513061 5234]	0.3493298618772 424	0.8714
	1,1,0.3	0.1918850362613 5484	0.2407096367532 3875	0.846611320972 4426 0.887451350688 9343 0.866200506687 1643	0.3313574500203 5333	0.8214
Phi-2	50,1,0.9	0.11206865202800 151	0.16469614391051 482	0.8340970277786 255 0.8877793550491 333 0.8596072793006 897	0.24880855334973 69	0.7642
	25,1,0.8	0.08959079242418 781	0.16469614391051 482	0.8340970277786 255 0.8877793550491 333 0.8596072793006 897	0.24880855334973 69	0.6416
	40,1,0.6	0.08959079242418 781	0.16469614391051 482	0.8340970277786 255 0.8877793550491 333 0.8596072793006 897	0.24880855334973 69	0.6416
	100,1,0.9	0.06634779432918 379	0.13412262386812 296	0.8001934885978 699 0.8386501073837 28 0.8188595175743 103	0.19478104415500 824	0.5714
Mistral	5,10,0.7	0.16289886520124 036	0.26387889788391 905	0.8444960713386 536 0.8742755055427 551 0.8584153056144 714	0.31484134692599 91	0.725
	20,15, 0.8	0.15854037102707 008	0.26455403371752 995	0.8439959883689 88 0.8706828355789 185 0.8564028739929 199	0.30603394971761 79	0.85
	50,20,1	0.17515414640534 432	0.30241677685686 67	0.8547066450119 019 0.8750926256179 81 0.8643683791160 583	0.36278138780242 59	0.87
	100,20,0.9	0.17496742843265 173	0.30015765172318 265	0.8544185757637 024 0.8743575811386 108 0.8638203144073 486	0.36831081766542 484	0.87

2. Write another discussion explaining the how the hyperparameters effect on the different metrics of LLaMA and Phi-2 (4-5 Lines).

The choice of hyperparameters affects the performance of LLaMA 2, Phi-2 and Mistral across different metrics. For LLaMA, beam size affects the outputs by picking the best output from several inferences, while temperature modulates randomness in outputs, I got better answers when Llama2 was in its preferred value of 10. Phi-2, on the other hand, shows top-k and beam size, with smaller values often leading to better results. For Mistral, I found Top-K of 50 or more helps improve the quality of the generated answer and the answers were more related to the instruction. Overall, finding the right balance between these hyperparameters is crucial for optimizing performance.