ARIA
APPLIED RESEARCH IN ACTION

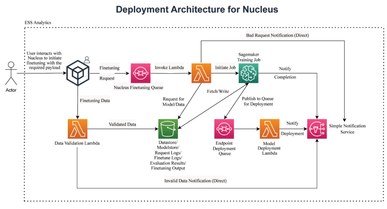
Enabling Large Language Models at Vanguard

Efficient Training and Usage of Domain Specific LLMs

Keerat Kaur Guliani

Rahul G. Krishnan
ACADEMIC SUPERVISOR

Jithin Pradeep, Ashish Bansal
INDUSTRY SUPERVISORS



Models / Benchmarks	ARC	HellaSwag	MMLU	TruthfulQA	Average Accuracy
Platypus2-70b-Instruct	71.84	87.94	70.48	62.26	73.13
Llama-2-70b-Instruct-v2	71.08	87.94	70.58	62.25	72.95
Falcon-40B-instruct	61.6	84.31	55.45	52.52	63.47
Llama-30b	61.26	84.73	58.47	42.27	61.68
GPT-Neo-Base-20b	45.65	74.03	29.92	34.51	46.03
GPT2-XL-1.5b	30.29	51.28	26.43	38.54	36.66
Charlie-Rhea	60.54	79.25	58.98	51.01	62.45

Table 1: Benchmarking Charlie-Rhea's performance against standard evaluation datasets. Charlie is Vanguard's own instruction-tuned LLM.

PROJECT SUMMARY

There are numerous practical applications at Vanguard which generate vast amounts of unstructured data holding crucial insights into client behaviour and needs. Recognizing this, there was a compelling requirement to adopt large language models (LLMs) on an enterprise-wide scale. Despite the availability of many open-source LLMs, they cannot be directly applied to Vanguard's internal use cases due to a lack of exposure to Vanguard-specific data and their time-consuming training and inference processes.

To this end, the project's primary objective was to develop a comprehensive framework that streamlines the training, fine-tuning, and inference processes of LLMs to better serve Vanguard's data science initiatives. The project kicked off by exploring various model designs, aimed at expediting both training and inference. Subsequently, Parameter Efficient Fine-tuning (PEFT)[1] techniques were leveraged to infuse Vanguard's internal knowledge into the LLM, which significantly reduced the training duration from several days to just a few hours, contingent on factors like model size and PEFT hyperparameters. Quantization techniques and fine-tuned backend configurations were implemented to reduce the storage requirements of the model without compromising its performance. This entire effort culminated in the introduction of Vanguard's very own large language model, called 'Charlie'. Table 1 demonstrates experimental results benchmarking Charlie's performance on different tasks, compared against other state-of-the-art models.

Charlie has been deployed as a package called 'Nucleus', which will be made available to different data science teams at Vanguard, enabling them to harness the power of LLMs for their specific needs. The Figure shows the deployment architecture which supports the release of Nucleus as an Amazon Sagemaker Python SDK. Nucleus promises to enhance data-driven insights and efficiency across the organization, making it an asset for Vanguard's future endeavours.

REFERENCES

[1] "PEFT." Available: https://huggingface.co/docs/peft/index

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing