LLM Implementation Inside of Software Development Work Environments

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jason Ton

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Rosanne Vrugtman , Department Computer Science

Technical Report

LLM Implementation Inside of Software Development Work Environments

CS4991 Capstone Report, 2024

Jason Ton
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA jht5qfn@virginia.edu

ABSTRACT

LLMs (Large Language Models), a rapidly growing technology in both engineering and the workplace, can serve as very useful tools for software development but it comes with many caveats, including data, privacy/ protection, resource requirements, accessibility, etc. During a summer internship with LinQuest, another intern and I experimented with implementing LLMs in a software development environment. The main model we attempted to implement was the BLOOM LLM with around 176 billion parameters. We had to do extensive research in the lab and close analysis of the resources available and what we needed to deploy the model. In my case this was not possible as the company resources allocated to my section's resources were not powerful enough to practically deploy the model. In terms of moving forward, my group worked on different solutions that could be more plausible, such as implementing models less computationally expensive.

1. INTRODUCTION

LLMs are a significant advancement in the field of artificial intelligence. These models take advantage of strong computational resources and advanced Machine Learning techniques to function. LLMs can understand, generate and interact with human speech and language with high levels of complexity like that of humans. They are trained in extensive articles of writing, including books, websites, texts, etc. In addition, they have evolved at an incredible pace and are only getting more advanced as more data and resources are becoming available. One of these reasons is that LLMs are more scalable. With more training data and resources, LLMs will become even more complex and intelligent.

When attempting to incorporate LLMs into a work environment, people often fail to notice the many complexities and obstacles that arise, including data privacy/protection, resource requirements, accessibility, etc. We have to remember that externally-hosted LLMs like ChatGPT are learning from training data. So, whenever we text them, this data goes into their database and is used to train their data. This means that giving them any proprietary information is illconceived, as they could steal ideas, methods, etc. A simple solution would be to host these LLMs within our own network, so the data does not leave our lab. However, it is incredibly

expensive and time-consuming to both set up and maintain.

2. RELATED WORKS

Many of the sources that discuss the obstacles of LLM deployment share the same concerns that LinQuest expressed. Regarding data privacy, Falconer (2023) noted: "Without robust anonymization or redaction measures in place, sensitive data becomes part of the model's training dataset, meaning that this data can potentially resurface later." Other issues, identified by Falconer and applicable in many fields include the inherent obstacles and risks that come with using externally hosted LLMs.

Riberio (2023) identified another challenge: the copious amounts of computing resources and time it takes to train and use LLMs. Like my lab situation, Riberio discusses the heavy need for specialized equipment like specific TPUs required for LLM training and computation and the priciness of these types of technologies. The daily cost of operation for ChatGPT is $700,000. Of course, smaller companies such as LinQuest would use an LLM as a supplementary tool, which would cost a great deal less, but running high tech TPUs and calculation tools all day would still cost quite a bit for smaller companies. In addition, LLMs are still relatively new to the public, making it hard to find advanced and experienced talent in the LLM field that knows how to manage and use this technology well enough to deploy and maintain it. Another concern is the cost of the power needed to keep the locally hosted LLM system running.

3. PROCESS DESIGN

The first step that my lab partner and I took was gathering knowledge on the capabilities of our lab. We went to our lab managers and determined the resources we had available. We figured out the Ram and GPUs the lab had available and looked at the computing power and capabilities of the available GPUs. We did research on our own to familiarize ourselves with the tools available. Our compute was 4 Tesla T4 GPUs. We investigated important stats such as the GPU Ram, Cores, TFOPs, etc. We looked into Nvidia's main website in order to learn more about the GPU and once we had our stats gathered, we worked on determining how powerful our GPUs actually were.

The other intern and I did not know exactly how strong the GPUs were, so we had to look into what kind of GPUs were needed in order to run and deploy the BLOOM mode, as discussed by Hotz (2022). We were trying to deploy 175B parameter BLOOM model and according to Hotz, that model needs about 8x A100 GPUS. We had no access to any A100s so we needed to figure out how strong our T4 GPUs were compared to the A100. We researched the stats of the A100 and learned that "the A Tesla T4 has 65 FP16 TFLOPS. A Tesla A100 has up to 312/624 FP16 TFLOPS" (Wehrens, 2023), meaning a Tesla A100 has roughly a difference of 5-10x computing power compared to a Tesla T4. In addition to GPU strength and TFLOPs, we also had to determine how much GPU Memory was needed, ultimately finding that we needed about 80GB of GPU Memory to host our BLOOM model.

We also investigated potential ways to make the most of what we had by trying parallel computing, potentially distributing computation load onto our 4 T4 GPUS at once to make the most of our resources. We also had to look into the amount of power our current T4 GPU would use and if it could be practical to run it 24/7 if the lab actually

found the model to be useful. We also wanted to look into potential solutions to efficiently work with GPU memory and figure out some way to make the most of the available resources. During this process there was a lot of compute analysis and looking into other powerful GPUs such as the RTX 2080 GPU. We wanted to see if there were potentially cheaper alternatives in case our T4 GPUs were not enough compute. We wanted to see if other, more cost-efficient GPUs could handle the BLOOM model. Harmon (2019) was a very helpful resource, as it provided visualizations showing how fast GPUs were at executing very popular machine learning algorithms essential to LLM functionality.

4. RESULTS

In the end we decided the available GPUs we had access to were not powerful enough. The requirements were 8 A100s and we only had access to 4 Tesla T4s. Because of this, we learned that our original goal of deploying the 175B parameter model was not practical and we reported to our management team that we lacked the computational power needed due to the lack of GPU Memory and TFLOPs. We did learn that our current GPUs had low memory consumption at around 70W, meaning that it would not be too expensive to run our GPUs 24/7; however we lacked everything else we needed in order to deploy.

As a result, we had to investigate potential alternatives to the 175B parameter BLOOM model. We were looking for smaller parameter versions of the BLOOM model, such as a few billion parameter versions. However after my partner and I investigated the power of the smaller models, we realized the T4 GPUs we had were not quite enough in order to host/deploy anything that would be useful to my team. My internship has now ended, but the team is looking into ways to potentially get company funds to upgrade our compute resources and practically deploy the 175B BLOOM model. They are currently working on getting a proof of concept for the 175B model of BLOOM and highlight the importance of keeping company information confidential and company secrets and data insourced.

5. CONCLUSION

The analysis of LLMs such as the BLOOM model within LinQuest's software development environment shows the gap between the computational resources that smaller companies have and the requirements to effectively deploy state-ofthe-art AI models. The experimentation revealed the harsh reality of the resourceintensive nature of such big models, highlighting how unrealistic it is for deployment of LLMs all over the world. While the lower power consumption of my available GPUs presented a sliver of hope in terms of cost, it did not compensate for the reality of computational capabilities needed for practical application.

My experience serves as a valuable lesson in the necessity of reality checking LLM deployment and the capabilities of most tech companies, as well as the importance of data privacy. Moving forward, companies must consider the importance of investing in advanced computing resources, or seeking alternative AI solutions that meet their goals. As AI technology continues to improve and become more readily available, the need for balance between integration and practical deployment will remain a critical point of consideration for organizations like LinQuest.

6. FUTURE WORK

My internship at LinQuest has ended but the Machine Learning team at LinQuest is

constantly looking to improve and integrate new and upcoming technologies. LinQuest is currently gathering and forming an argument to advocate for LinQuest's acquisition of more Machine Learning technology. Like many other small companies, LinQuest recognizes the importance of AI and its significance for the future. As a result, many companies are scrambling to acquire LLM-capable technology and to recruit people knowledgeable in the LLM field.

REFERENCES

Falconer, S. (2023, October 23). Privacy in the Age of Generative AI - Stack Overflow. Stackoverflow.blog. https://stackoverflow.blog/2023/10/23/p r ivacy-in-the-age-of-generative-ai/

Ribeiro, D. (2023, November 15). THE UNSPOKEN CHALLENGES OF LARGE LANGUAGE MODELS [Review of THE UNSPOKEN CHALLENGES OF LARGE LANGUAGE MODELS]. Deeper Insights. https://deeperinsights.com/aiblog/the-unspoken-challenges-of-largelanguagemodels#:~:text=Deploying%20LLMs%20involves%20multiple%20network,user%20experience%20and%20processing%20efficiency.

Hotz, H. (2022, August 16). *Let's bloom with BigScience's new AI model*. Medium. https://towardsdatascience.com/letsbloom-with-bigscience-s-new-ai-model-803b1a0d677

Nvidia Tesla T4 GPU. (n.d.). https://www.itcreations.com/nvidiagpu/nvidia-tesla-t4-gpu#:~:text=The%20NVIDIA%20Tesla%20T4%20enterprise,wide%20range%20of%20modern%20applications.

Wehrens, O. (2023, December 9). *OpenAI whisper benchmark nvidia tesla T4 / A100*. owehrens.com. https://owehrens.com/openai-whisperbenchmark-on-nvidia-tesla-t4-a100/

Harmon, W. (2019b, October 3). *Nvidia Tesla T4 AI inferencing GPU benchmarks and Review*. ServeTheHome. https://www.servethehome.com/nvidiatesla-t4-ai-inferencing-gpu-review/4/