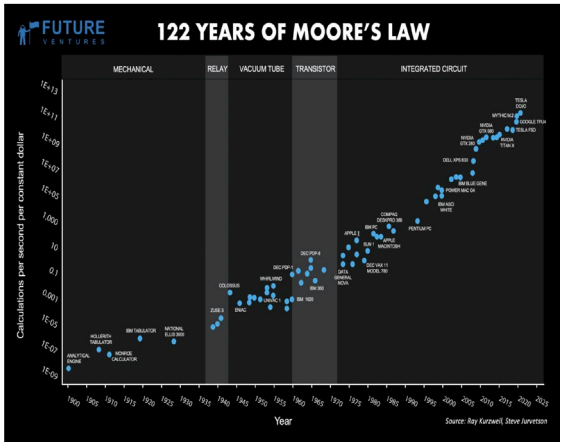
GPTforALL

Aditya Thurvas Senthil Kumar
Heena Chandak
Shreya Ajay Kale

Motivation

Large language models (LLMs) like GPT-3 have indeed showcased remarkable natural language processing capabilities, as evidenced by the performance of ChatGPT. However, their training necessitates significant computational resources, typically only accessible to major tech companies, posing a barrier to broader adoption. Even fine-tuning, which tailors pretrained models to specific tasks and data, demands substantial infrastructure and expertise. There exists a pressing need to democratize access to custom LLMs, especially for niche applications, empowering small businesses and individuals to harness the potential of this technology effectively.

Release	Model	Size	Paper
2019	GPT-2	1.5B	Language Models are Unsupervised Multitask Learners
2020	GPT-3	175B	Language Models are Few-Shot Learners
2021	Gopher	280B	Scaling Language Models: Methods, Analysis & Insights from Training Gopher
2022	PaLM	540B	PaLM: Scaling Language Modeling with Pathways
2022	Chinchilla	70B	Training Compute-Optimal Large Language Models
2022	OPT	175B	OPT: Open Pre-trained Transformer Language Models
2022	BLOOM	176B	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
2022	Galactica	120B	Galactica: A Large Language Model for Science
2023	LLaMA	65B	LLaMA: Open and Efficient Foundation Language Models

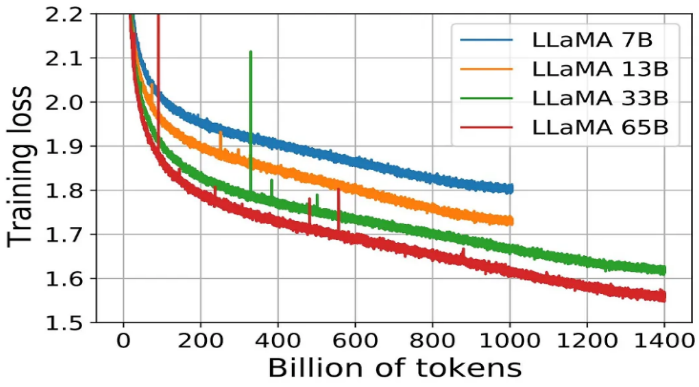


Training Costs

Estimating the cost of training large language models requires careful consideration of three primary factors inherent to any machine learning algorithm: data, compute resources, and expertise

Data

For instance, LLaMA utilized a training dataset containing an impressive 1.4 trillion tokens, totalling a massive 4.6 terabytes in size.



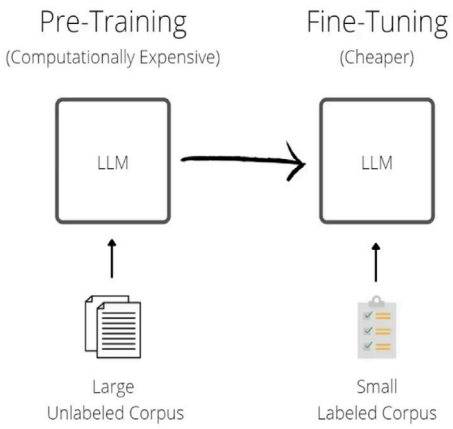
Training loss over tokens for LLaMA models. Figure 1 from LLaMA paper

Compute

The training phase constitutes a significant portion of the budget allocated for Large Language Model (LLM) development. Training such expansive language models necessitates substantial resources, primarily relying on high-performance Graphics Processing Units (GPUs) renowned for their robust parallel processing capabilities.

Notably, NVIDIA continually releases powerful GPUs annually, each priced in the hundreds of thousands of dollars. The cost of employing cloud computing services for training these models can escalate dramatically, often reaching several million dollars, particularly when factoring in the need for iterations through diverse configurations.

Motivation for Fine-Tuning



In the domain of artificial intelligence, Large Language Models (LLMs) such as GPT-3.5 reign as powerful tools for understanding and generating human language. Fine-tuning, a critical technique, stands at the forefront for organizations striving to produce personalized and contextually relevant content. This process allows the customization of general language understanding attained during pre-training, aligning the model with specific applications and use cases. However, it's important to note that the initial pre-training of these large language models on extensive and diverse text data enriches their comprehension of general language nuances, grammar, and context. Despite this foundational understanding, challenges arise due to their vast parameter count, which can limit a comprehensive grasp of technical jargon and effective generalization, particularly when faced with data limitations.

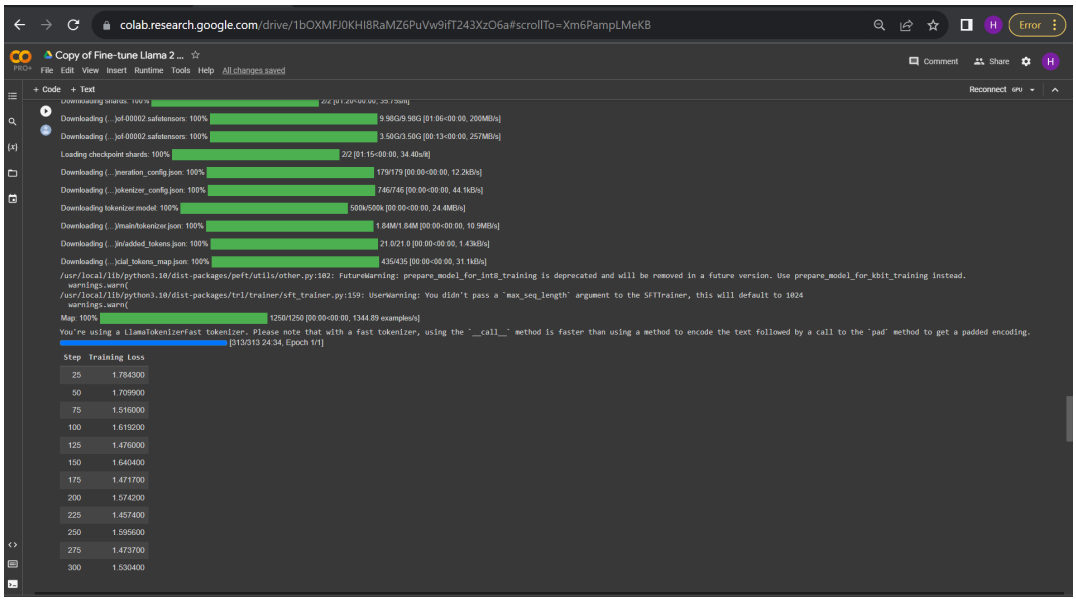
Additionally, zero-shot learning with a small dataset poses a significant challenge, as sparse examples may not adequately capture the diversity and variations within different classes, hindering accurate classification. Furthermore, the responsible deployment of AI in this domain necessitates careful considerations regarding data privacy, bias mitigation, and ethical implications to ensure beneficial and equitable use of this transformative technology.

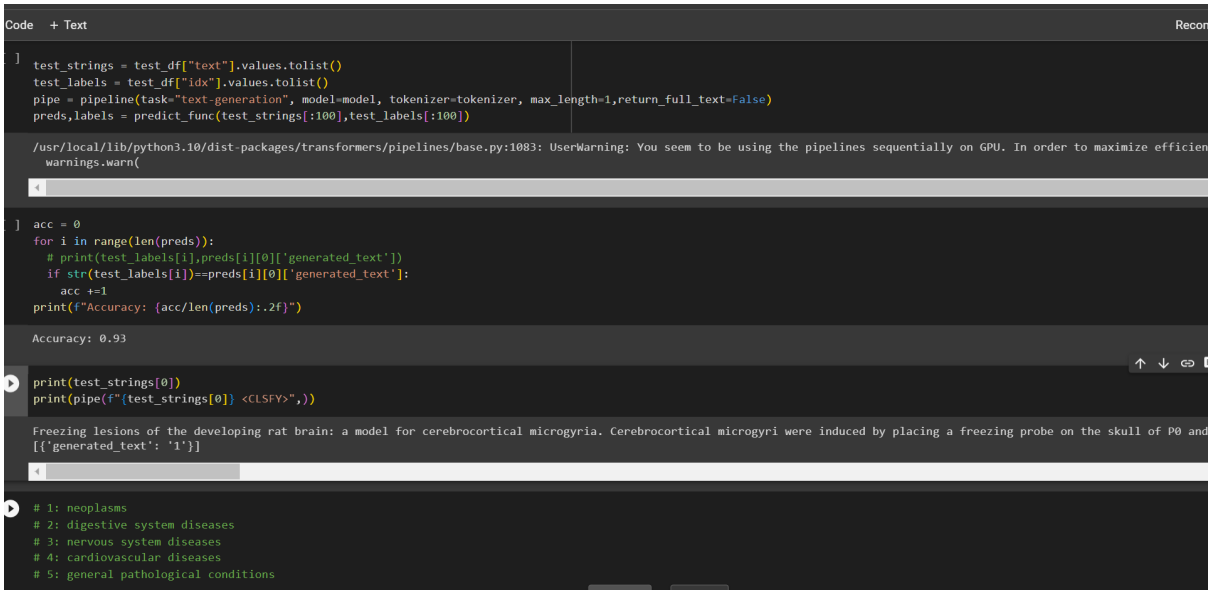
Methods

The proposed **GPTforAll** allows on low resource device training of customized LLMs with an intuitive UI, removing dependencies on expensive infrastructure. It leverages autoML for efficient finetuning, with estimated 10-100x efficiency gains over standard approaches. GPTforAll builds on top of LLaMA, an open source 7B parameter model optimized for efficiency using 8-bit quantization. For medical text classification, GPTforAll enables training specialized models on 14k labeled abstracts across 5 disease categories and evaluating on 14k unlabeled samples. Model Environment - Google Colab with Nvidia T4 GPUs, achieving a total training time of approximately 24 minutes for a single epoch with 1,250 datapoints.

Results

In our study, we leveraged 1,250 labelled abstracts for training, encompassing five distinct disease classes: Digestive, Cardiovascular, Neoplasms, Nervous System, and General Pathological conditions. The training was conducted in the Google Colab environment, utilizing Nvidia T4 GPUs. Impressively, the total training time amounted to approximately 24 minutes for a single epoch, demonstrating the efficiency of our approach with this dataset size. Furthermore, our model showcased a high accuracy rate of 93%, underscoring its effectiveness in medical text classification.





Future Work

GPTforAll models can be evaluated on domain-specific benchmarks to quantify efficiency gains and customized performance compared to generic pretrained models. Additional research into responsible and ethical development of democratized LLMs is needed as the technology proliferates.

References

LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS https://arxiv.org/pdf/2106.09685.pdf

LLaMA: Open and Efficient Foundation Language Models https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/

PEFT: Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware https://huggingface.co/blog/peft

https://www.kaggle.com/datasets/chaitanyakck/medical-text