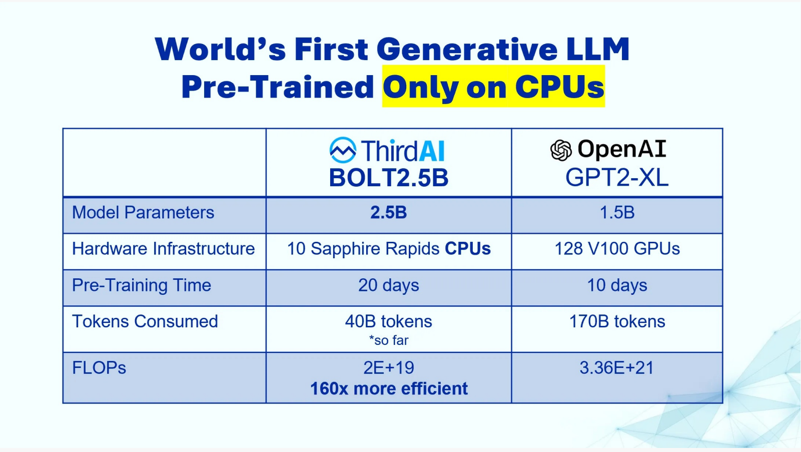
EINPRESSWIRE

ThirdAI Unveils BOLT2.5B: The World’s First Generative LLM Trained Only on CPUs

Introducing the World’s First Generative LLM Pre-Trained Only on CPUs. The model is fine-tunable on commodity CPUs in a matter of minutes.



World’s First Generative LLM
Pre-Trained Only on CPUs

	ThirdAI BOLT2.5B	OpenAI GPT2-XL
Model Parameters	**2.5B**	1.5B
Hardware Infrastructure	10 Sapphire Rapids **CPUs**	128 V100 GPUs
Pre-Training Time	20 days	10 days
Tokens Consumed	40B tokens *so far	170B tokens
FLOPs	2E+19 **160x more efficient**	3.36E+21

Bolt comparison with GPT2

HOUSTON, TEXAS, UNITED STATES, September 27, 2023 / EINPresswire.com/ -- ThirdAI, an AI company specializing in efficient & affordable AI training, has hit a significant landmark in the Generative AI space. Today, they proudly announce the launch of BOLT2.5B, a 2.5-billion parameter Generative Large Language Model (LLM) that has been trained exclusively on CPUs. A game-changing move away from traditional GPU-focused models, BOLT2.5B is set to redefine what's possible with CPU-trained AI.

> “
> We are at the dawn of a new GenAI era, where the capabilities of commodity compute meet the demands of AI. BOLT2.5B is not just a model; it's a testament to the untapped power of CPUs in the AI realm.”
>
> *Anshumali Shrivastava*

Key Insights into BOLT2.5B:

Rapid Pre-training: In just 20 days, ThirdAI achieved pre-training of the model on 10 Sapphire Rapids CPUs, under the Intel Disruptor program, processing a staggering 2 billion tokens each day, totaling around 40 billion tokens.

Remarkable Performance: Preliminary evaluations show BOLT2.5B's capabilities rivaling OpenAI’s GPT-2 XL model, which was trained using 128 V100s over 10 days, processing a phenomenal 170 billion tokens.

Unlocking the Power of CPUs for GenAI:

Train Your Own: With ThirdAI’s innovative approach, pre-training a GenAI from scratch is now astonishingly simple. Users can harness cloud CPUs or data centers to craft their custom BOLT GenAI models with ease.

Fine-tuning Redefined: BOLT2.5B's 'dynamic sparse' architecture allows even older desktops to fine-tune the model, with rates reaching up to 50 tokens per second. An illustration provided by ThirdAI showcases the model being fine-tuned on a Shakespearean corpus, achieving one epoch of fine-tuning in just 20 minutes on a 2014 dual socket Intel(R) Xeon(R) CPU E5–2680 v3 server

Swift Inference: Designed with CPUs in mind, BOLT2.5B delivers rapid inference, producing tokens at a rate of 20 per second without the need for specialized treatments like quantization or pruning.

"We are at the dawn of a new GenAI era, where the capabilities of commodity computing rise to meet the demands of AI. BOLT2.5B is not just a model; it's a testament to the untapped power of CPUs in the AI realm. With this launch, we invite everyone to experience this transformative shift," says Anshumali Shrivastava, Founder & CEO of ThirdAI.

ThirdAI has hosted BOLT2.5B on Hugging Face and provided a Github notebook for any user wishing to experiment. With only a requirement of 10GB of RAM for inference and 20GB for fine-tuning, there's no longer a need to wait for GPUs to develop and fine-tune a custom LLM.

About ThirdAI: ThirdAI is a Houston-based AI company that aims to make deep learning technology more affordable and attainable for all. It builds and deploys large language models on CPUs without the need for configurations or expensive GPUs. ThirdAI's models reduce latency, minimize training time, and increase accuracy by retraining only specific parameters when needed, rather than updating the entire AI model. ThirdAI enables companies to train LLMs on their own data with existing hardware, lowering the cost and energy consumption associated with AI training.

For more information visit our website at https://www.thirdai.com/bolt-25b/

Vinod Iyengar
ThirdAI
+ 19087648639
email us here
Visit us on social media:
Twitter
LinkedIn
YouTube

Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.