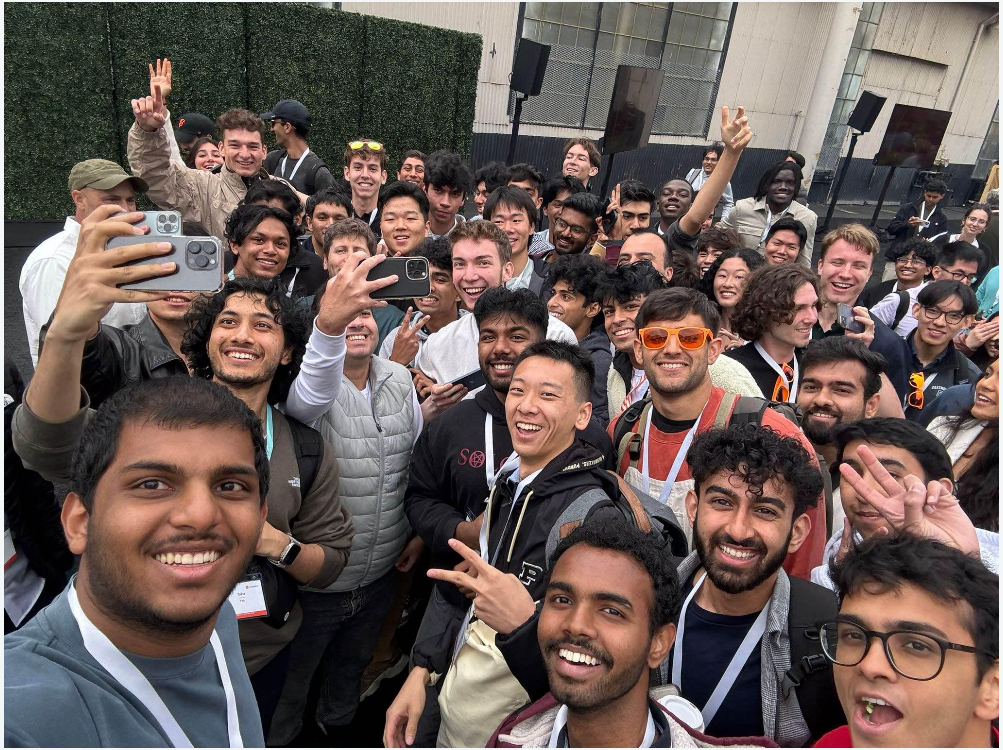
Computing for ChatGPT, LLM

By Joshua Goodman and Dean Mah

Motivation



Description of the technology: What is AI Supercomputing?

More formally known as : **AI Supercomputing**
Leverages high-performance GPUs/TPUs, parallel computing, and distributed frameworks to process massive datasets and train advanced AI models.

1. Thousands of cores for parallel computations.
2. High memory bandwidth for managing large datasets like NVLink and InfiniBand.
3. Enhanced by AI-specific hardware accelerators like TPUs (Google) and AMD Instinct GPUs.
4. Often paired with high-capacity storage systems for handling petabyte-scale datasets.

Impact of this technology

1. Forcing technology to shift in a new direction (Ex. Chip Manufacturing)
2. Increased access to immense computing power
3. Boosts AI-driven innovation in startups, big tech, and cloud providers
    - Ai as a service
4. Increased power consumption
    - Drives demand for renewable energy solutions to mitigate environmental impact.
5. Increased cooling resource requirement
6. Increased prices for consumer hardware (GPU’s)

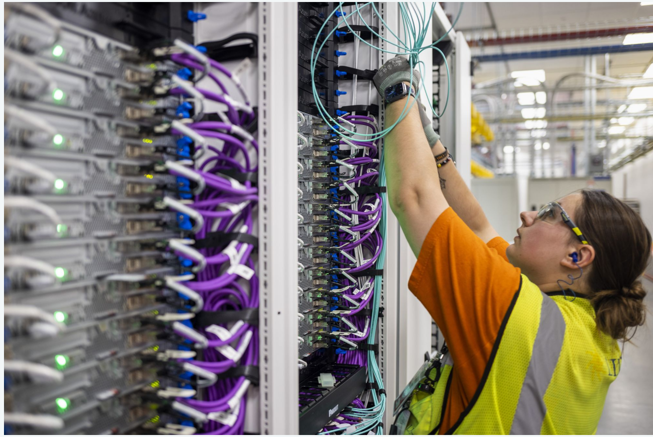
Strengths and limitations

Strengths:

- Exceptional speed for parallel tasks.
- Scalable to handle massive datasets and models.
- Drives innovation across industries.

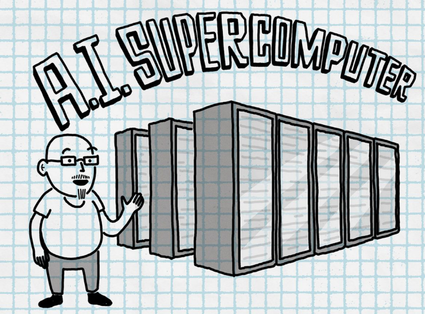
Limitations:

- High costs for hardware and energy consumption.
- Complex setups requiring skilled professionals.
- Environmental impact due to energy usage.
- Network Bandwidth



Why is this technology relevant and how is it being used?

1. OpenAi uses a A.I. Supercomputer built by microsoft for ChatGPT.
    a. Includes over 285,000 AMD InfiniBand connected CPU cores
    b. 10,000 NVIDIA V100 Tensor Core GPUs
    c. 400 gigabits per second of network connectivity for each GPU server
        i. GPT-3 cost 700,000k per day to run.
        ii. Cost of training GPT-4 was more than $100 million.
            1. GPT-4 has estimated more than 1 trillion parameters



Real world applications

Model Training: Leveraging GPUs and TPUs for parallelized operations during large-scale neural network training.

Simulations: Using GPT-like models for simulating language interactions in complex systems, such as robotics or autonomous systems.

Rendering and Graphics: Powering CGI in movies, video game graphics, and AR/VR development.

Cryptocurrency Mining: Solving cryptographic puzzles in blockchain systems.

Performance statistics

Model Training:

- GPT-3: 175 billion parameters, trained on 800 GB of text over weeks using thousands of NVIDIA GPUs.
- Energy Cost: ~10s of megawatts for large-scale model training.

Inference Latency:

- Real-time responses generated in milliseconds.

Data Handling:

- Processes petabytes of data with distributed storage systems.

Technology maturity: Where are we at now?

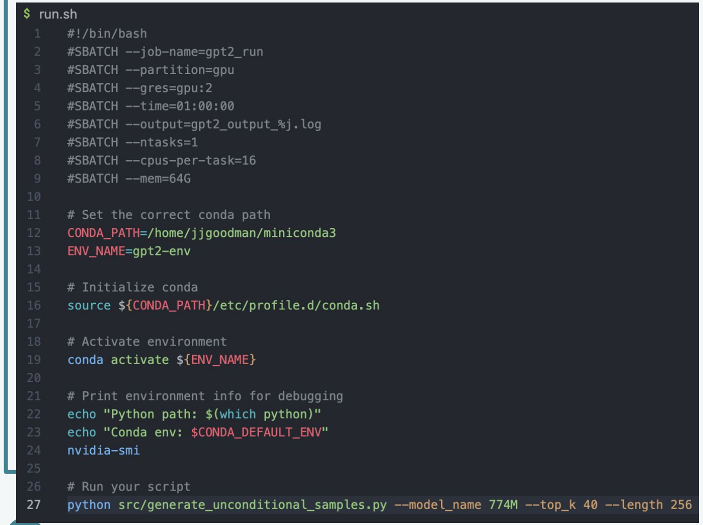
Infant, but argued to be “Production Ready”
Production Ready - Powers over 1 million companies world wide, daily.
Infant - Very poorly optimized, and only 2 years old.

- Likely a 9/9 for the Technology Readiness Level (TRL)
    - Already being sold on the market and tested by users
    - Still an infant in our opinion due to being so new

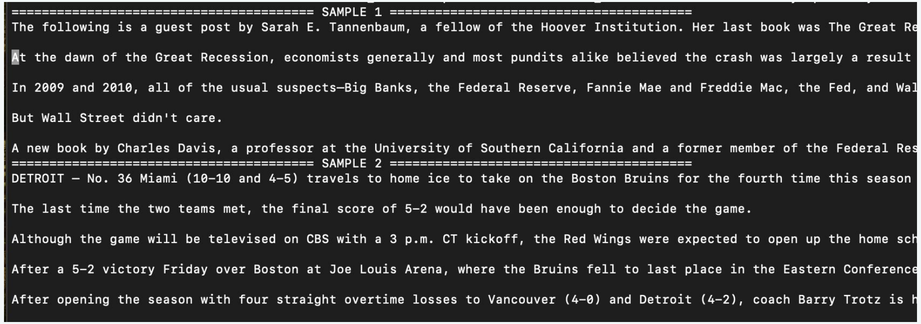
DEMO

Running GPT2 on CPP HPC environment:

Backup video link: https://youtu.be/AGwYorbxvgg



```bash
$ run.sh
#!/bin/bash
#SBATCH --job-name=gpt2_run
#SBATCH --partition=gpu
#SBATCH --gres=gpu:2
#SBATCH --time=01:00:00
#SBATCH --output=gpt2_output_%j.log
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=64G

# Set the correct conda path
CONDA_PATH=/home/jjgoodman/miniconda3
ENV_NAME=gpt2-env

# Initialize conda
source ${CONDA_PATH}/etc/profile.d/conda.sh

# Activate environment
conda activate ${ENV_NAME}

# Print environment info for debugging
echo "Python path: $(which python)"
echo "Conda env: $CONDA_DEFAULT_ENV"
nvidia-smi

# Run your script
python src/generate_unconditional_samples.py --model_name 774M --top_k 40 --length 256
```



```
======================================== SAMPLE 1 ========================================
The following is a guest post by Sarah E. Tannenbaum, a fellow of the Hoover Institution. Her last book was The Great Re

At the dawn of the Great Recession, economists generally and most pundits alike believed the crash was largely a result 

In 2009 and 2010, all of the usual suspects—Big Banks, the Federal Reserve, Fannie Mae and Freddie Mac, the Fed, and Wal

But Wall Street didn't care.

A new book by Charles Davis, a professor at the University of Southern California and a former member of the Federal Res
======================================== SAMPLE 2 ========================================
DETROIT — No. 36 Miami (10-10 and 4-5) travels to home ice to take on the Boston Bruins for the fourth time this season 

The last time the two teams met, the final score of 5-2 would have been enough to decide the game.

Although the game will be televised on CBS with a 3 p.m. CT kickoff, the Red Wings were expected to open up the home sch

After a 5-2 victory Friday over Boston at Joe Louis Arena, where the Bruins fell to last place in the Eastern Conference

After opening the season with four straight overtime losses to Vancouver (4-0) and Detroit (4-2), coach Barry Trotz is h
```