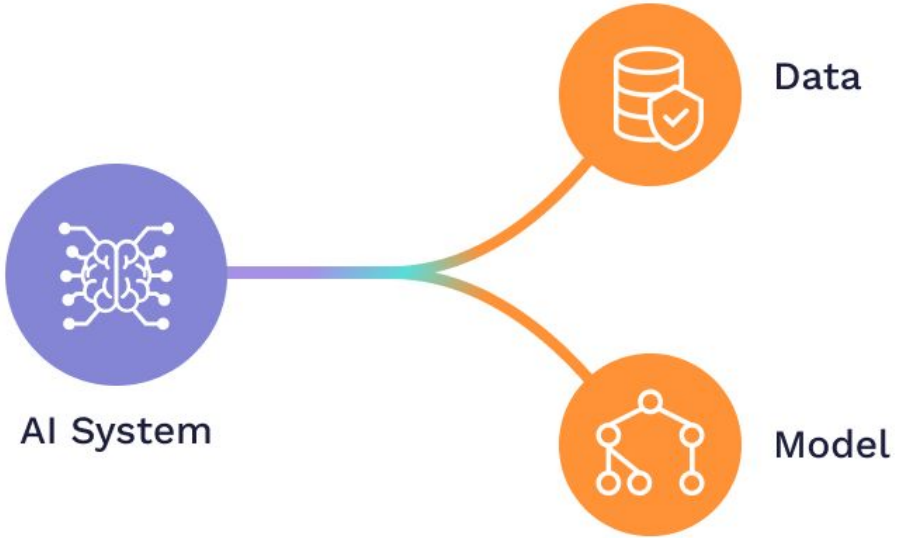
Machine Learning System for Sustainable and Affordable AI

Presenter: Leon Le

Harker Research Symposium 2024

Three Pillars of Artificial Intelligence

- **Data**: The raw material used to train and evaluate machine learning models.
- **Model**: A mathematical representation that learns from data for predictions or decisions.
- **System**: The infrastructure supporting the development and operation of machine learning models.



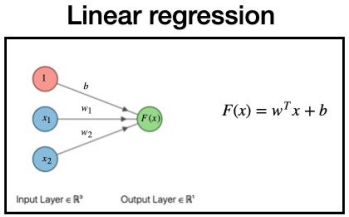
Data: The Oil for Artificial Intelligence

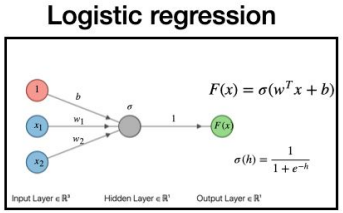
- **Social Media**: User interactions on platforms like Facebook, Twitter, and Instagram.
- **Web and Search Data**: Data from search engines, website logs, and browsing activity.
- **Mobile Devices**: Data from smartphones, including location, app usage, and communication logs.
- **Online Transactions**: Data from e-commerce, online banking, and digital interactions.
- **Internet of Things (IoT) Devices**: Data from sensors and smart devices.

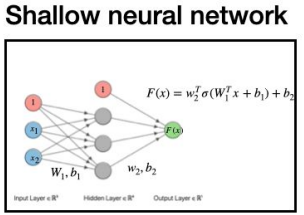
https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and-applications/

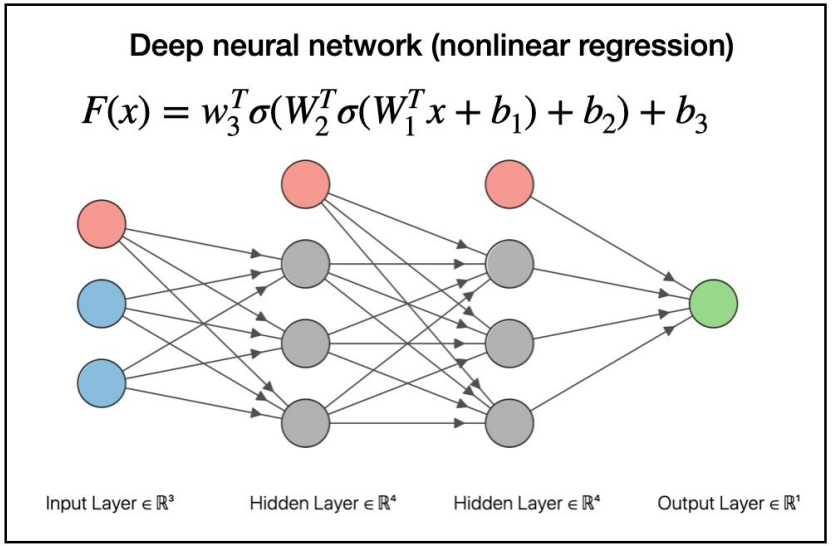
Model: The Brain of Artificial Intelligence

- **Supervised Learning**: Learns from labeled data to predict outcomes for new data.
  - Linear Regression,
  - Logistic Regression
  - Neural Networks
- **Unsupervised Learning**: Explores unlabeled data to find hidden patterns or intrinsic structures.
- **Reinforcement Learning**: Learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward.





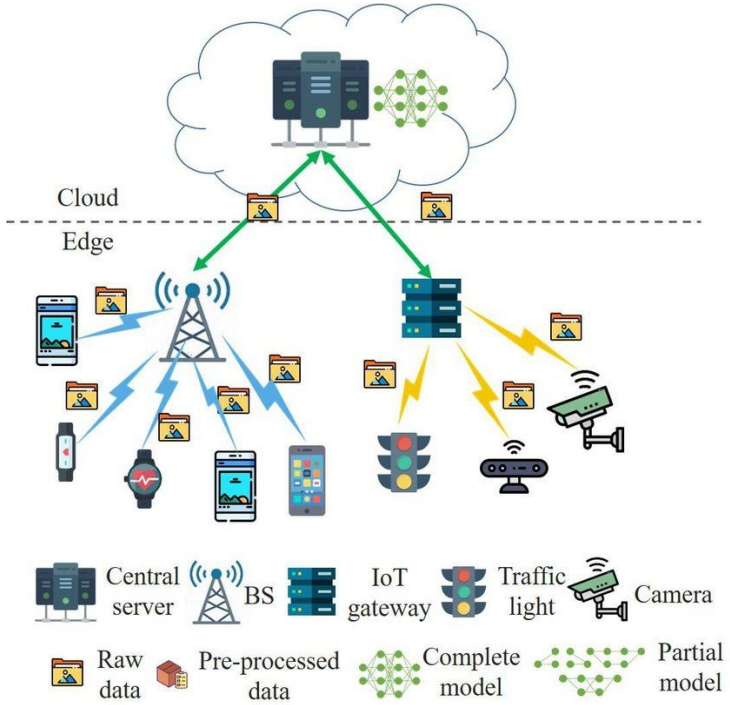




https://joshuagoings.com/2020/05/05/neural-network/

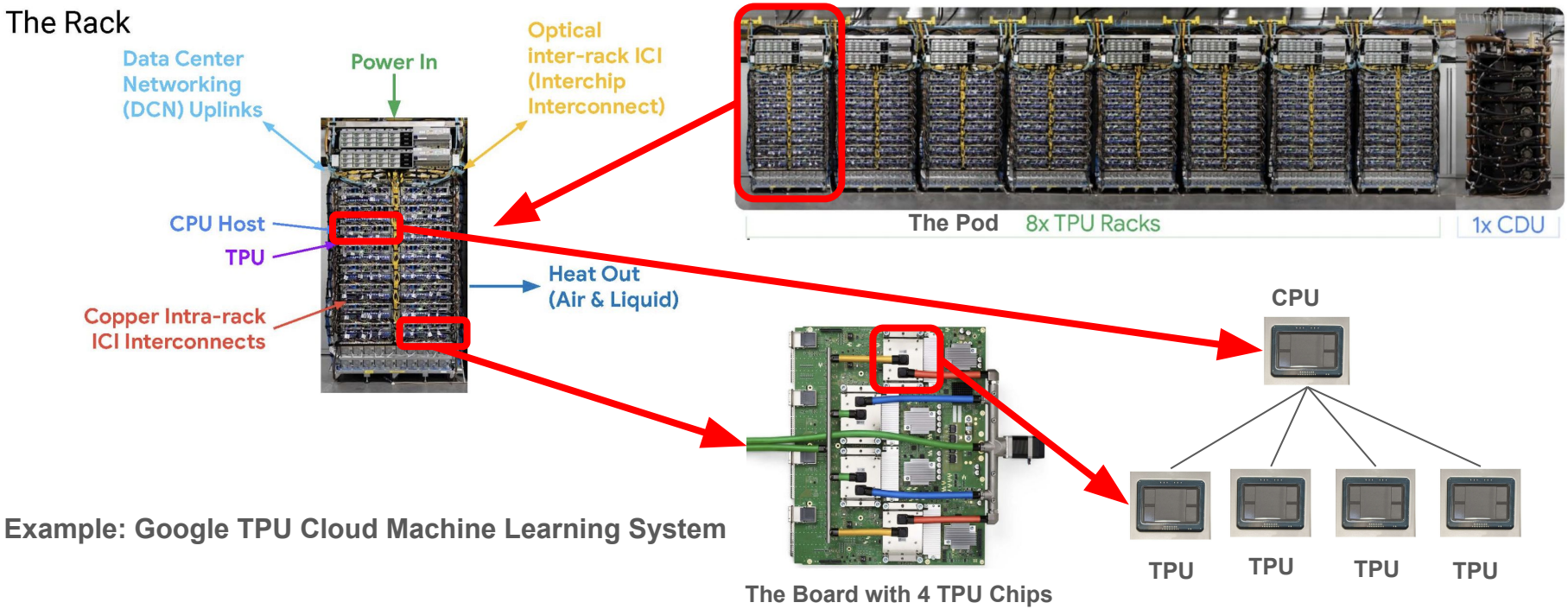
System: The Engine of Artificial Intelligence

- Cloud (Data Center) Machine Learning Systems allow users to build, train, and deploy machine learning models using cloud resources.
- Mobile (Edge) Machine Learning Systems allows machine learning models to run directly on mobile devices, enabling on-device processing, low latency, privacy, offline functionality, and optimized resource use.



https://ieeexplore.ieee.org/document/9596610

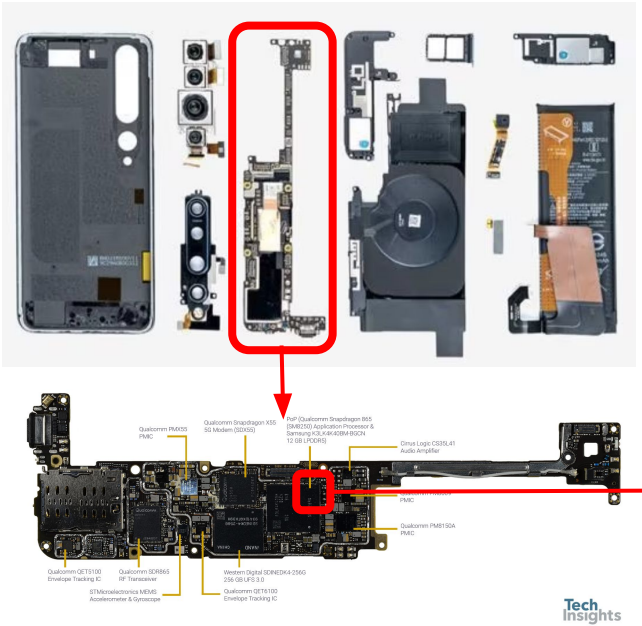
Cloud Machine Learning Systems

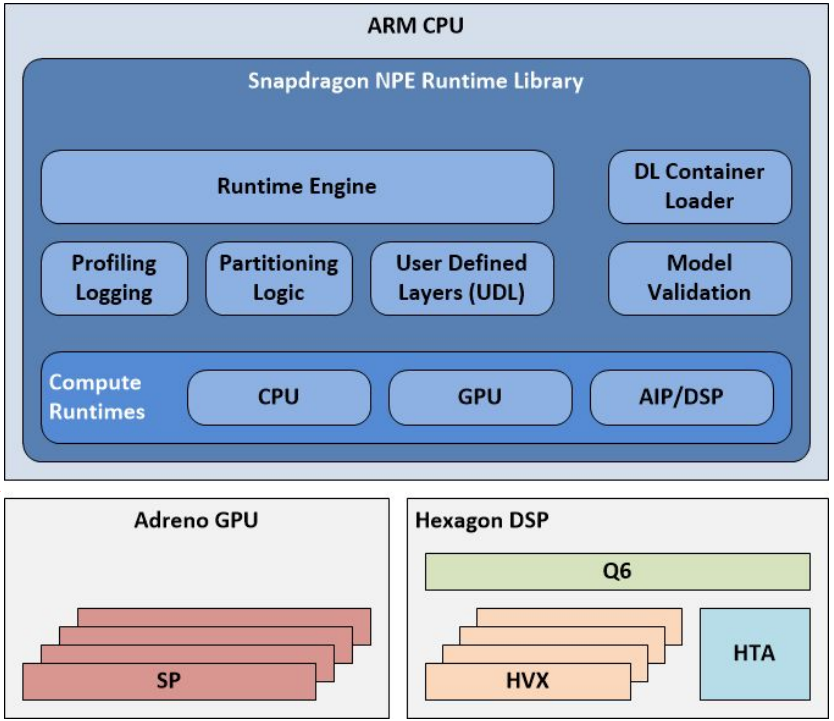


Example: Google TPU Cloud Machine Learning System

https://www.servethehome.com/google-details-tpuv4-and-its-crazy-optically-reconfigurable-ai-network/

Mobile (Edge) Machine Learning Systems



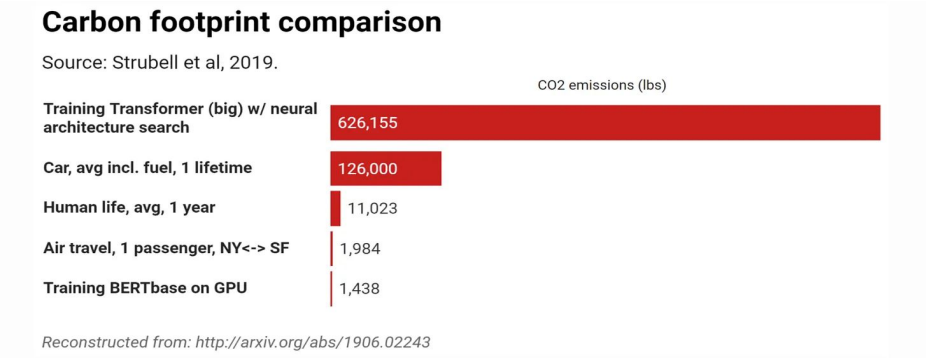


Example: Android Mobile Machine Learning System

https://developer.qualcomm.com/sites/default/files/docs/snpe/snapdragon_npe_runtime.html

Sustainability for Artificial Intelligence

- Concerns are growing over the environmental impact of complex ML models' high computational needs.
- The carbon footprint of ML systems is increasing due to the significant energy used in training and operating these models.
- Analysing and enhancing ML system performance reduces energy usage and supports sustainable AI.

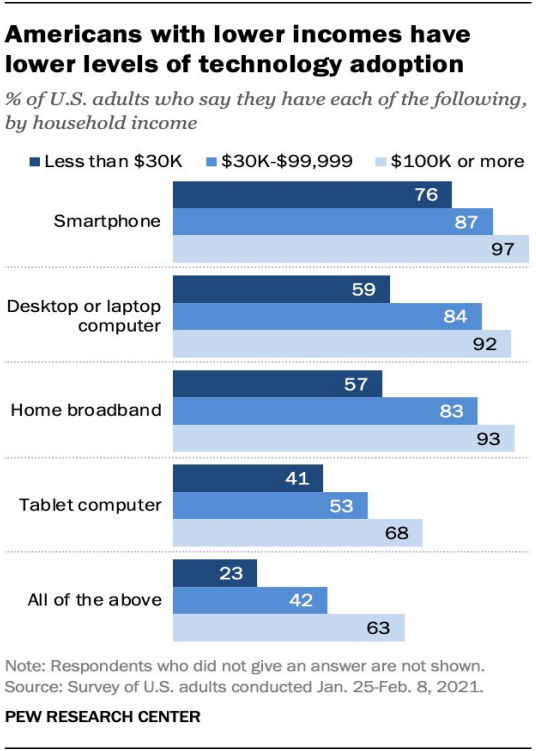


Google has shown that AI contributes to 10–15% of their overall electricity usage, which was 18.3 terawatt hours in 2021. That would mean that Google's AI burns around 2.3 terawatt hours annually, nearly double the amount of electricity required to power the London underground annually.

https://medium.com/@audaciatech/sustainable-ai-3-tools-to-measure-the-environmental-impact-of-ml-solutions-1762c88626cc

Affordability for Artificial Intelligence

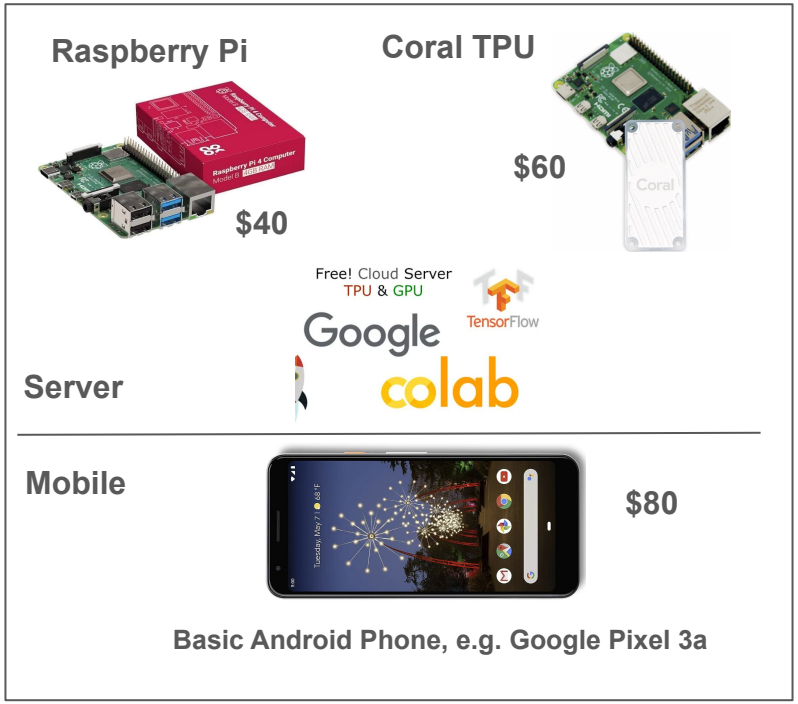
- High costs limit access to the latest AI technologies for individuals and small organizations.
- Creates a digital divide between those who can afford advanced AI and those who cannot.
- Restricts the potential user base and hampers inclusivity in the tech world.
- Impacts equity by preventing equal access to AI-powered solutions in sectors like healthcare and education.



https://www.pewresearch.org/short-reads/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/

Proposal: A Low-cost Machine Learning System for Sustainable and Affordable Artificial Intelligence

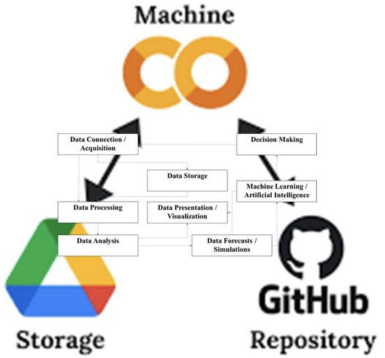
- We propose an alternative cost effective full-stack machine learning system.
- Improve the performance of the ML system to reduce dependency on cloud.
- Develop the applications of the ML system to increase affordability, especially for underprivileged communities.



A full-stack ML System that cost less than $200

Experiment 1: Performance Profile for Meta Llama 2 LLM in Google Cloud Collab

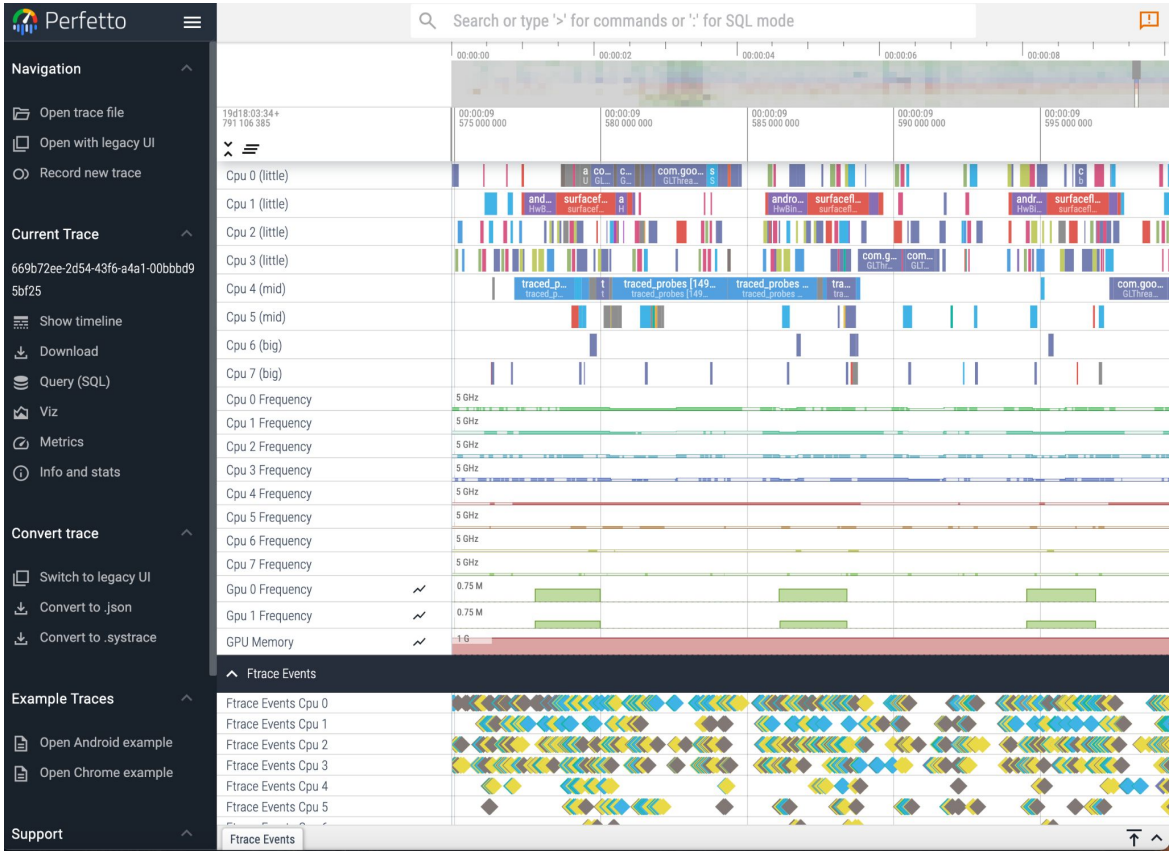
```python
from torch.profiler import profile, record_function, ProfilerActivity

# Ignore warnings
logging.set_verbosity(logging.CRITICAL)

with profile(activities=[ProfilerActivity.CPU, ProfilerActivity.CUDA]) as prof:
    # Run text generation pipeline with our next model
    prompt = "What is a large language model?"
    pipe = pipeline(task="text-generation", model=model, tokenizer=tokenizer, max_length=200)
    result = pipe(f"<s>[INST] {prompt} [/INST]")

print(result[0]['generated_text'])
print(prof.key_averages().table(sort_by="cuda_time_total", row_limit=10))
```



Name	Self CPU %	Self CPU	CPU total %	CPU total	CPU time avg	Self CUDA	Self CUDA %	CUDA total	CUDA time avg
MatMul4Bit	9.74%	4.129s	21.80%	9.240s	222.964us	25.575s	55.18%	44.822s	1.082ms
void kDequantizeBlockwise<__half, 512, 64, 8, 2>(flo...	0.00%	0.000us	0.00%	0.000us	0.000us	25.575s	55.18%	25.575s	617.147us
aten::matmul	1.21%	514.888ms	10.04%	4.255s	79.587us	0.000us	0.00%	19.354s	361.986us
aten::mm	3.80%	1.611s	6.42%	2.720s	65.349us	17.751s	38.30%	18.575s	446.256us
aten::linear	0.61%	260.090ms	8.27%	3.507s	84.245us	0.000us	0.00%	17.788s	427.341us
turing_fp16_s1688gemm_fp16_256x64_sliced1x2_ldg8_f2f...	0.00%	0.000us	0.00%	0.000us	0.000us	8.057s	17.38%	8.057s	559.516us
turing_fp16_s1688gemm_fp16_128x128_ldg8_f2f_tn	0.00%	0.000us	0.00%	0.000us	0.000us	5.678s	12.25%	5.678s	376.530us
cudaLaunchKernel	12.61%	5.344s	12.61%	5.344s	16.639us	3.840s	8.28%	3.840s	11.956us
turing_fp16_s1688gemm_fp16_128x64_sliced1x2_ldg8_f2f...	0.00%	0.000us	0.00%	0.000us	0.000us	3.211s	6.93%	3.211s	316.039us
aten::mul	1.90%	804.919ms	3.23%	1.367s	25.227us	735.754ms	1.59%	1.355s	24.994us

Self CPU time total: 42.388s
Self CUDA time total: 46.350s

Experiment 2: Performance Profile for running Apps on Mobile Phone



Summary

- We proposed a low-cost machine learning system for sustainable and affordable artificial intelligence.
  - Examined the three pillars of AI, focusing on machine learning systems.
  - Reviewed sustainability and affordability issues of artificial intelligence.
  - Proposed a cost-effective machine learning system (below $200).
  - Demonstrated how to do performance analysis for our system.
- Our research is in its early stages, and we welcome feedback to explore further opportunities.