

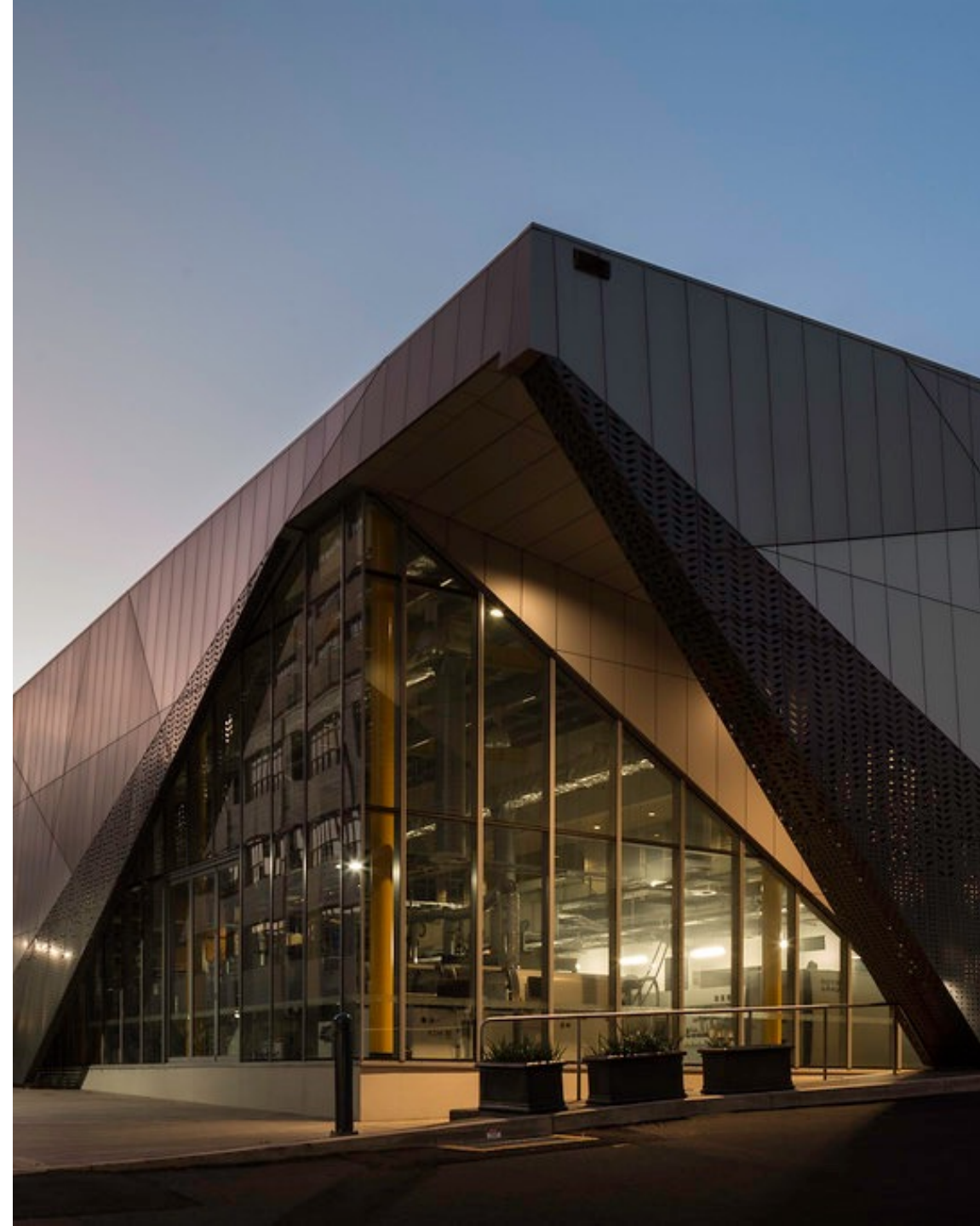


A Tailored Large Language Model for Science

Team members: Tong Xie, Qingyuan Linghu, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu

Mentor: Darren Tan, Bharatkumar

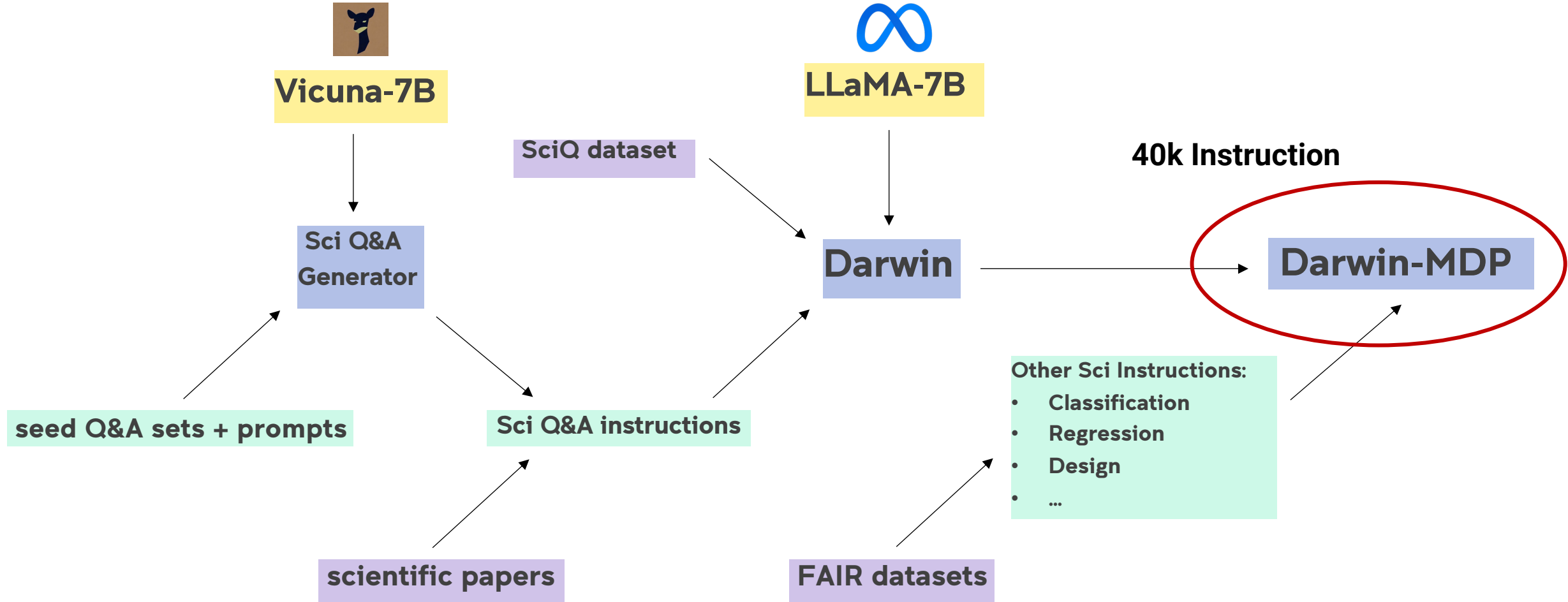
NCI OpenHackthon 2023





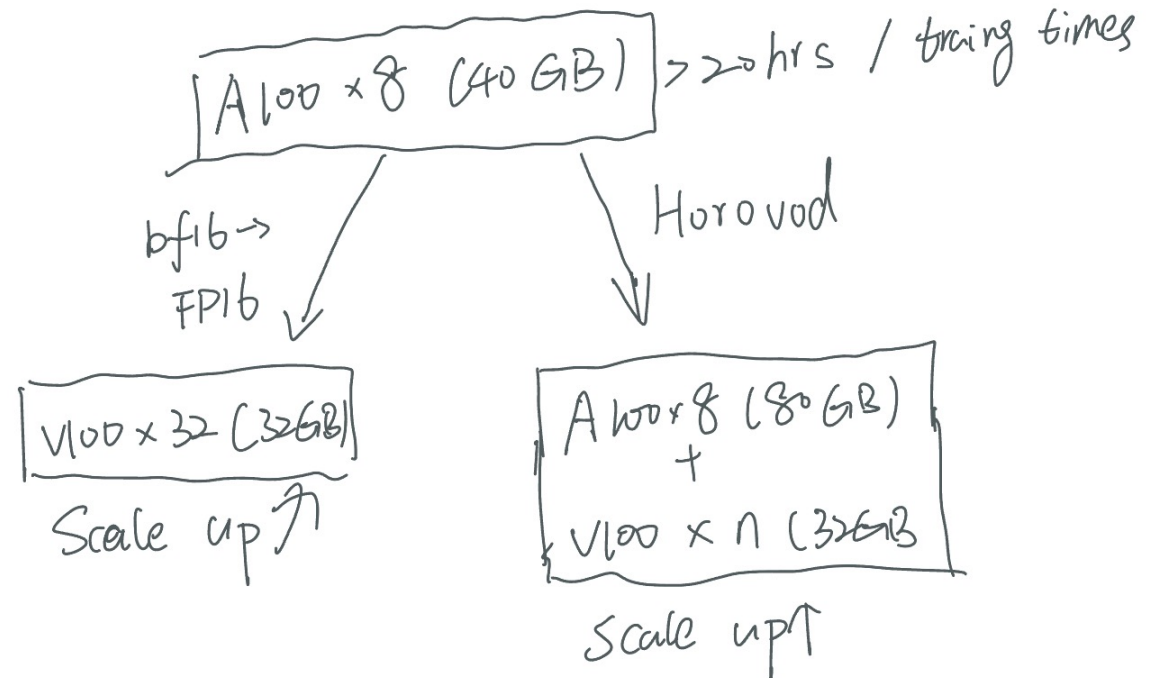
Open AI is not open anymore, expensive to do research on it

Transformer architecture



Evolution and Strategy

- What was your goal coming here?
Faster speed in training, inference
- What was your initial strategy?
We already use Pytorch and A100
- How did this strategy change?
A100 + V100 -> only A100 or V100
Horovod -> DeepSpeed
AdamW -> Sophia



Possible Solution

Lower the Precision to FP16 to make it can run on V100 multinode

















Two new tools

DeepSpeed: an easy-to-use deep learning optimization software suite that enables unprecedented scale and speed for DL Training and Inference.

Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training

Results

fp16 vs bf16

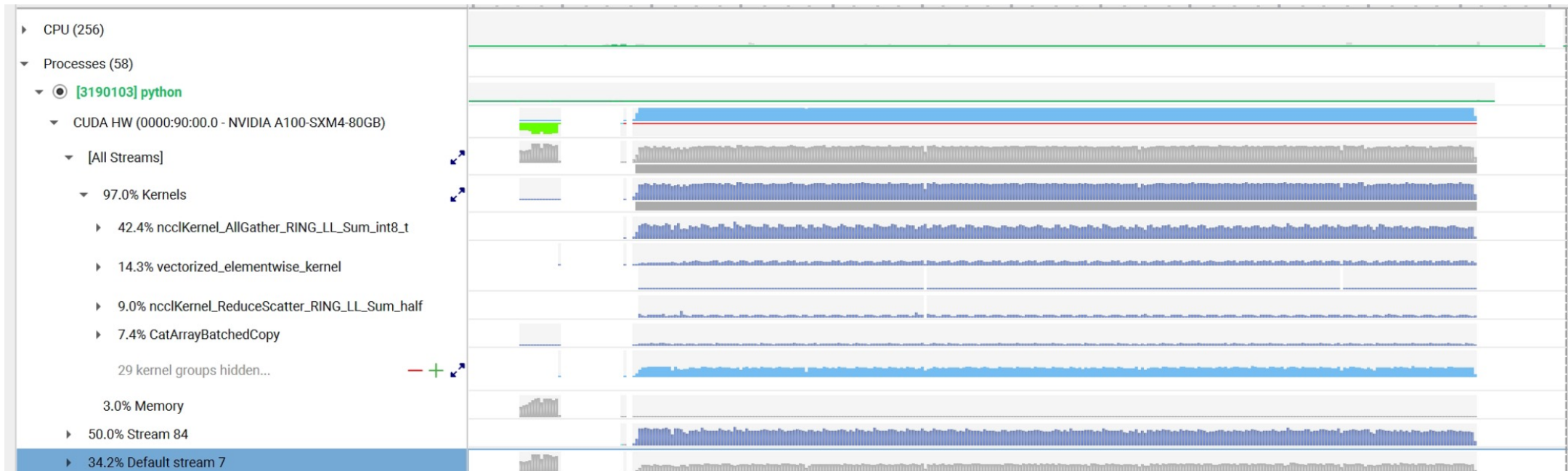
•   8gpu_fp16	 Finished	Add notes	plumyi	2d ago	7m 45s	-
•   8gpu_bf16	 Finished	Add notes	plumyi	2d ago	7m 47s	-
•   4gpu-bf16_full_eval	 Finished	Add notes	plumyi	2d ago	44m 23s	-
•   4gpu-fp16	 Finished	Add notes	plumyi	2d ago	15m 33s	-
•   4gpu-bf16	 Finished	Add notes	plumyi	 2d ago	15m 18s	-

Profiler Output

8 * A100 (80GB) 7min 47s

3k Instruction

Detected the bottleneck: communication between GPUs



Possible Solution 3k Instruction

8 * A100 (80GB) with DeepSpeed 2min30s(3x faster!)



Energy Saving

INPUTS	
# CPU Cores	128
# GPUs (A100)	1
Application Speedup	3.0x
Node Replacement	24.0x

GPU NODE POWER SAVINGS			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (W)	26,400	6,500	19,900
Networking Power (W)	1,114	93	1,022
Total Power (W)	27,514	6,593	20,922

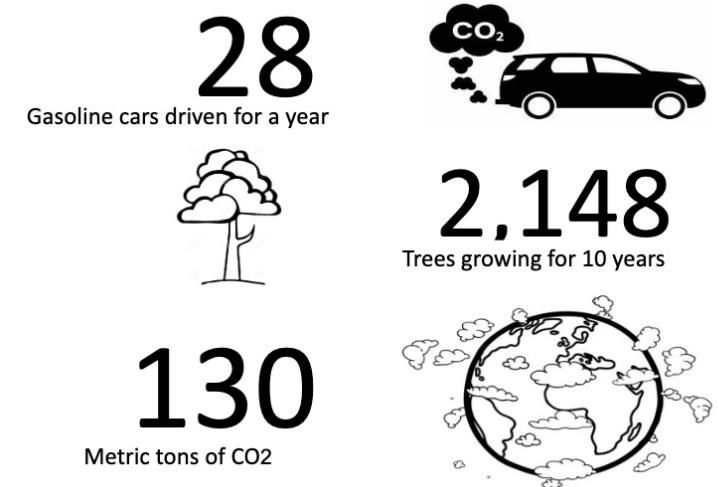
Node Power efficiency	4.2x
------------------------------	-------------

ANNUAL ENERGY SAVINGS PER GPU NODE			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (kWh/year)	231,264	56,940	174,324
Networking Power (kWh/year)	9,763	814	8,949
Total Power (kWh/year)	241,027	57,754	183,273

\$/kWh	\$ 0.34
Annual Cost Savings	\$ 62,312.92
3-year Cost Savings	\$ 186,938.75

Metric Tons of CO2	130
Gasoline Cars Driven for 1 year	28
Seedlings Trees grown for 10 years	2,148

[\(source: Link\)](#)



Doing

8 * A100 (80GB) with Sophia

```
{ 'loss': 4.2581, 'learning_rate': 0.000870927952239912, 'epoch': 0.92}
31% ██████████ | 212/690 [02:13<05:00, 1.59it/s]
{ 'loss': 5.0545, 'learning_rate': 0.0008716324053443783, 'epoch': 0.92}
31% ██████████ | 213/690 [02:14<04:59, 1.59it/s]
{ 'loss': 7.6416, 'learning_rate': 0.0008723334470402997, 'epoch': 0.92}
{ 'loss': 4.129, 'learning_rate': 0.0008730311102089144, 'epoch': 0.93}
31% ██████████ | 215/690 [02:15<04:39, 1.70it/s]
{ 'loss': 8.2452, 'learning_rate': 0.0008733786849912526, 'epoch': 0.93}
{ 'loss': 0.0, 'learning_rate': 0.0008733786849912526, 'epoch': 0.94}
31% ██████████ | 217/690 [02:16<03:46, 2.08it/s]
{ 'loss': 0.0, 'learning_rate': 0.0008733786849912526, 'epoch': 0.94}
```

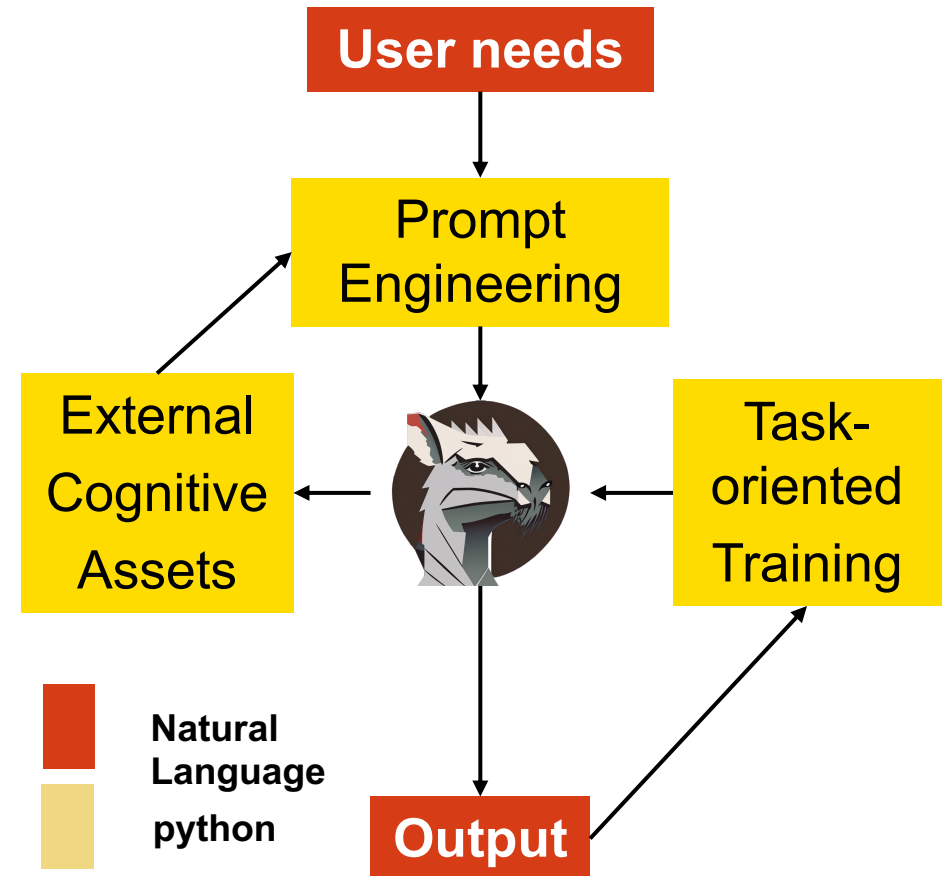
LLaMA with V100 on GADI

Potential Work

LangChain with Darwin

```
question = "What NFL team won the Super Bowl in the year Justin Bieber was born?"
```

```
llama.cpp: loading model from /srv/scratch/z5293104/llama-7b-hf/ggml-model-f16.bin
llama_model_load_internal: format      = ggjt v1 (pre #1405)
llama_model_load_internal: n_vocab    = 32000
llama_model_load_internal: n_ctx      = 512
llama_model_load_internal: n_embd     = 4096
llama_model_load_internal: n_mult     = 256
llama_model_load_internal: n_head     = 32
llama_model_load_internal: n_layer    = 32
llama_model_load_internal: n_rot      = 128
llama_model_load_internal: ftype      = 1 (mostly F16)
llama_model_load_internal: n_ff       = 11008
llama_model_load_internal: n_parts    = 1
llama_model_load_internal: model size = 7B
llama_model_load_internal: ggml ctx size = 0.07 MB
llama_model_load_internal: mem required = 14645.09 MB (+ 1026.00 MB per state)
.
llama_init_from_file: kv self size = 256.00 MB
AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM
_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 0 | SSE3 = 1 | VSX = 0 |
We can use algebra and substitution. First we will find what year the Superbowl was held that
matched Justin's birthday. To do so, we will add 22+1960 = 1982 (this is because he was born
in March of that year). Now, let's work backwards. Substitute this date into what we already k
```

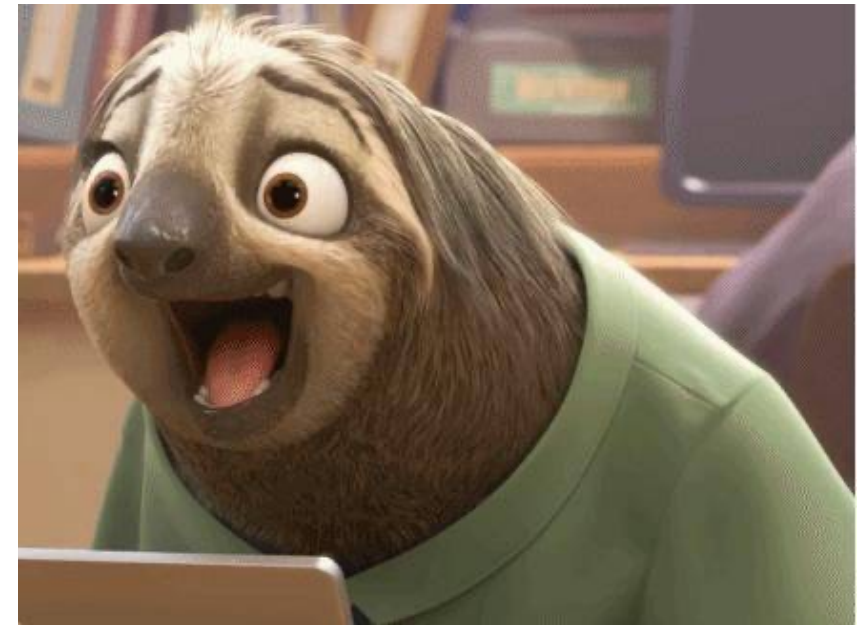


Potential Work

LangChain with Darwin with CPU

```
question = "What NFL team won the Super Bowl in the year Justin Bieber was born?"
```

```
llama.cpp: loading model from /srv/scratch/z5293104/llama-7b-hf/ggml-model-f16.bin
llama_model_load_internal: format      = ggjt v1 (pre #1405)
llama_model_load_internal: n_vocab    = 32000
llama_model_load_internal: n_ctx      = 512
llama_model_load_internal: n_embd    = 4096
llama_model_load_internal: n_mult    = 256
llama_model_load_internal: n_head     = 32
llama_model_load_internal: n_layer    = 32
llama_model_load_internal: n_rot      = 128
llama_model_load_internal: ftype      = 1 (mostly F16)
llama_model_load_internal: n_ff       = 11008
llama_model_load_internal: n_parts    = 1
llama_model_load_internal: model size = 7B
llama_model_load_internal: ggml ctx size = 0.07 MB
llama_model_load_internal: mem required = 14645.09 MB (+ 1026.00 MB per state)
.
llama_init_from_file: kv self size = 256.00 MB
AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM
_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 0 | SSE3 = 1 | VSX = 0 |
We can use algebra and substitution. First we will find what year the Superbowl was held that
matched Justin's birthday. To do so, we will add 22+1960 = 1982 (this is because he was born
in March of that year). Now, let's work backwards. Substitute this date into what we already k
```



Potential Work

LangChain with CuBLAST Acceleration

```
ggml_init_cublas: found 2 CUDA devices:
  Device 0: Tesla V100-SXM2-32GB
  Device 1: Tesla V100-SXM2-32GB
llama.cpp: loading model from /srv/scratch/z5293104/llama-7b-hf/ggml-model-f16.bin
llama_model_load_internal: format      = ggjt v1 (pre #1405)
llama_model_load_internal: n_vocab    = 32000
llama_model_load_internal: n_ctx      = 512
llama_model_load_internal: n_embd     = 4096
llama_model_load_internal: n_mult     = 256
llama_model_load_internal: n_head     = 32
llama_model_load_internal: n_layer    = 32
llama_model_load_internal: n_rot      = 128
llama_model_load_internal: ftype      = 1 (mostly F16)
llama_model_load_internal: n_ff       = 11008
llama_model_load_internal: n_parts    = 1
llama_model_load_internal: model size = 7B
llama_model_load_internal: ggml ctx size = 0.07 MB
llama_model_load_internal: using CUDA for GPU acceleration
ggml_cuda_set_main_device: using device 0 (Tesla V100-SXM2-32GB) as main device
llama_model_load_internal: mem required = 2292.09 MB (+ 1026.00 MB per state)
llama_model_load_internal: allocating batch_size x 1 MB = 512 MB VRAM for the scratch buffer
llama_model_load_internal: offloading 32 layers to GPU
llama_model_load_internal: offloading output layer to GPU
llama_model_load_internal: total VRAM used: 12865 MB
.....
llama_init_from_file: kv self size = 256.00 MB
AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM
_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 1 | SSE3 = 1 | VSX = 0 |
The Superbowl has been played each year since 1
```

2 V100 GPU
=
64 AMD CPU

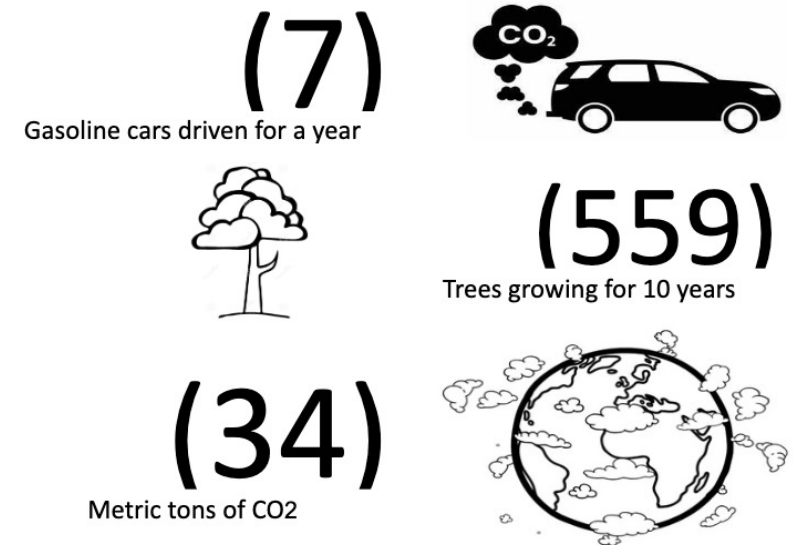
Energy Efficiency

INPUTS	
# CPU Cores	64
# GPUs (A100)	4
Application Speedup	1.0x
Node Replacement	1.0x

GPU NODE POWER SAVINGS			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (W)	1,100	6,500	-5,400
Networking Power (W)	46	93	-46
Total Power (W)	1,146	6,593	-5,446
Node Power efficiency	0.2x		

ANNUAL ENERGY SAVINGS PER GPU NODE			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (kWh/year)	9,636	56,940	(47,304)
Networking Power (kWh/year)	407	814	(407)
Total Power (kWh/year)	10,043	57,754	(47,711)
\$/kWh	\$ 0.34		
Annual Cost Savings	\$ (16,221.67)		
3-year Cost Savings	\$ (48,665.00)		
Metric Tons of CO2	(34)		
Gasoline Cars Driven for 1 year	(7)		
Seedlings Trees grown for 10 years	(559)		

[\(source: Link\)](#)



❖ Tailored GPT for natural science



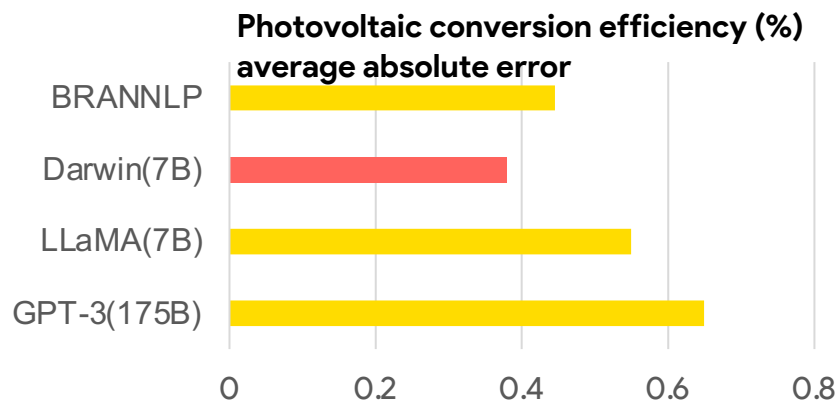
❖ Powerful materials design capabilities

Darwin (7B) > Fine Tuned GPT-3 (175B)
Multi-task training >> Single-task training
One Darwin, All Tasks and Open Source!

Case:





 **HOPV15**



❖ Better scientific answering than GPT-4


What are the key properties of graphitic carbon nitride (g-C₃N₄) that make it an attractive material for optoelectronic applications?

 Graphitic carbon nitride (g-C₃N₄) possesses a large bandgap, high thermal and chemical stability, and excellent photocatalytic properties, making it an attractive material for optoelectronic applications such as solar cells, sensors and photocatalyst,

 Graphitic carbon nitride (g-C₃N₄) exhibits excellent semiconductor properties, high thermal and chemical stability, and strong optical absorption in the visible light range. Its bandgap can be tuned for specific applications. Moreover, it's non-toxic, abundant, and eco-friendly, making it ideal for sustainable optoelectronic applications.

What's the bandgap of Ag₂CsSb₂I₃?

 1.96

 As of my knowledge cutoff in September 2021, there isn't specific information available on the bandgap of Ag₂CsSb₂I₃. The bandgap of a material is a very specific property,I recommend checking the most recent scientific literature for updates or conducting an experiment if possible.

PROBLEMS 4

OUTPUT

DEBUG CONSOLE

TERMINAL

JUPYTER

python - tong + v [] [] ... ^ X

User: What phenomenon makes global winds blow northeast to southwest or the reverse in the northern hemisphere and northwest to southeast or the reverse in the southern hemisphere?

#Response: coriolis effect

User: Boron only occurs naturally in compounds with what element?

#Response: oxygen

User: What is phase of CdTe?

#Response: multi-phase

User: What is crystal st of given alloy? $Ag_{0.05}Zr_{0.95}$

#Response: bcc

User: Given composition, write its band gap. $\rightarrow ZnSn_3$

#Response: 0.0

User: How is the water stability of given stability at room temperature and atmospheric pressure? $\rightarrow Zn_4O(1,4\text{-benzenedicarboxylate-(CH}_3)_2)_3$

#Response: It has low water stability with low confidence.

User: Design a donor with acceptor PC71BM, PCE= 10 for a organic solar cell

#Response: CN1C(=O)c2csc(-c3cc4c(s3)-c3sccc3C4(C)C)c2C1=O

User: what is the power conversion efficiency of organic solar cells with donor : COC(=O)c1cc2csc(-c3cccs3)c2s1, and Acceptors: PC61BM?

Conclusion

- Was this worth it?

Yes!

- Will you continue development?
 - Keep optimizing our model

More A100 will make my life much more more more easier!

Thanks 🙏





Thanks

Team members: Tong Xie, Qingyuan Linghu, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu
Mentor: Darren Tan, Bharatkumar

NCI OpenHackthon 2023

