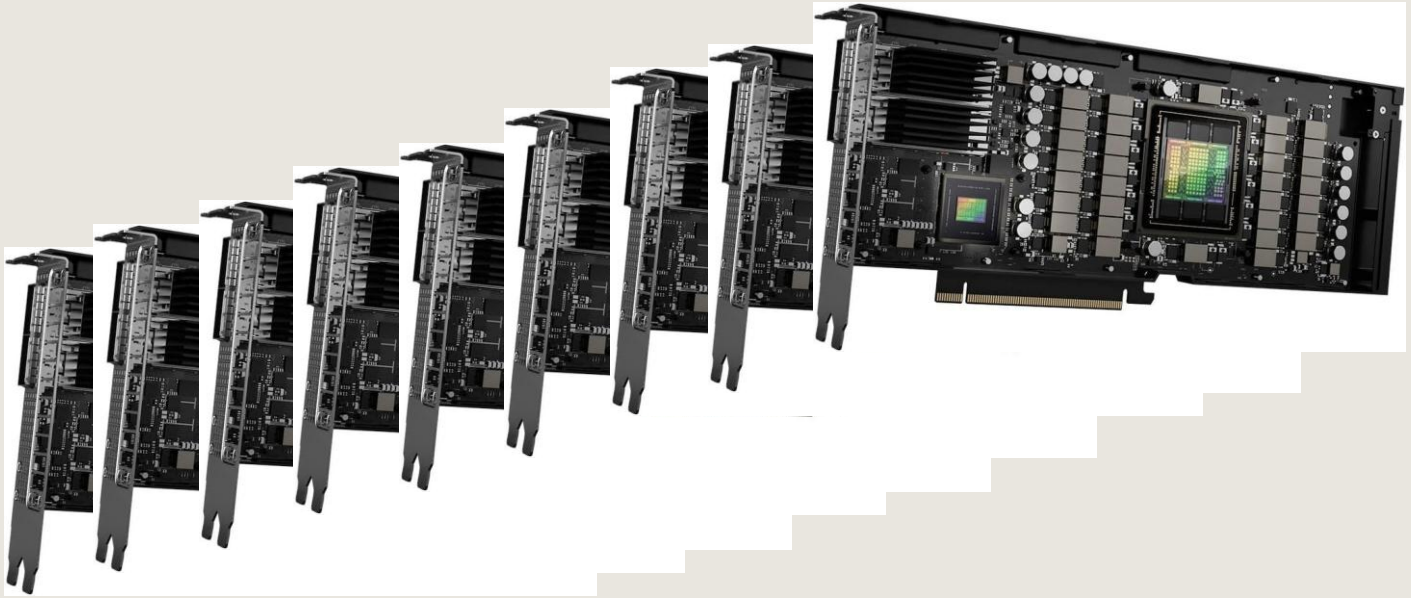
NanoGPTQ

Philipp Sepin

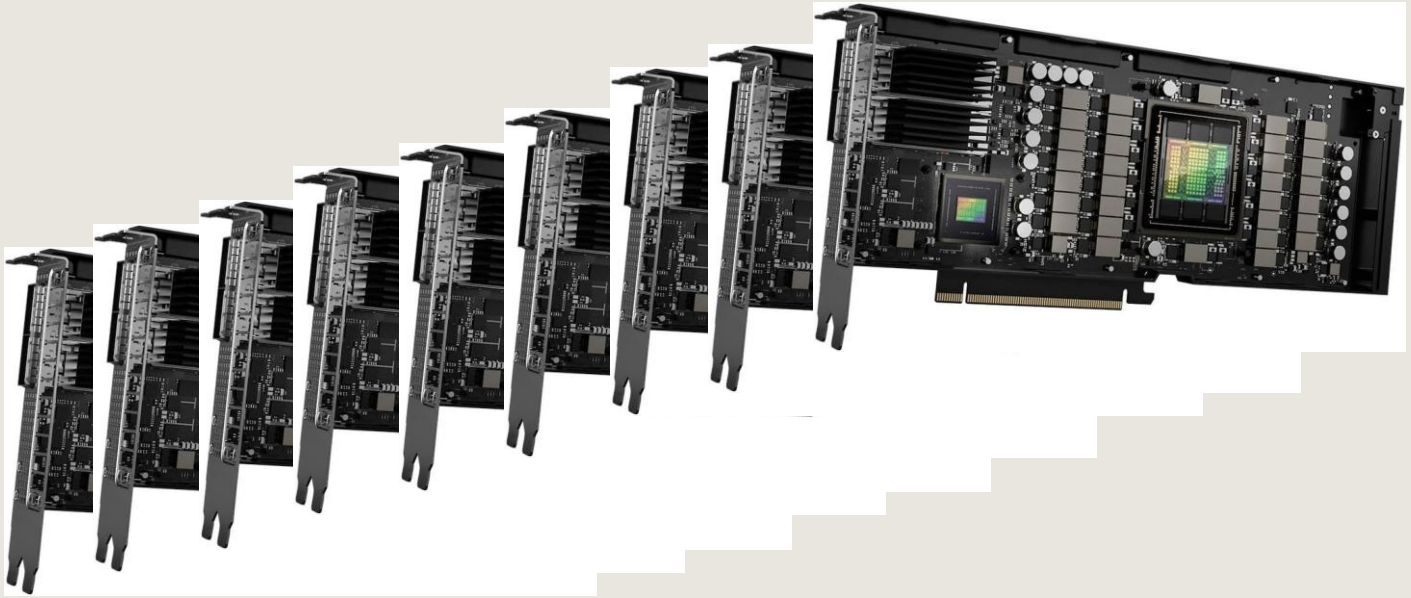
LARGE LANGUAGE MODELS

- Very powerful but very large
- GPT-3 has 175 billion weights



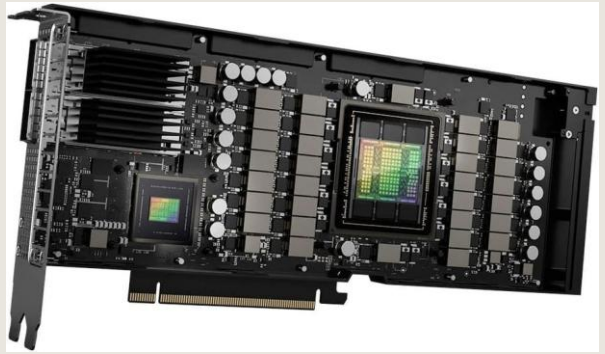
LARGE LANGUAGE MODELS

- Very powerful but very large
- GPT-3 has 175 billion weights
- Takes up around 700 GB in GPU memory with float32 weights
- 9 H100 GPUs needed



QUANTIZATION

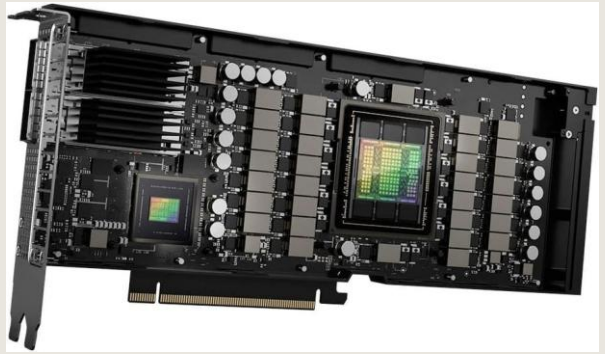
- Changing float32 weights to int8 or int4
- By naively rounding it to the next integer
- Or more sophisticated algorithms like GPTQ



QUANTIZATION

- Changing float32 weights to int8 or int4
- By naively rounding it to the next integer
- Or more sophisticated algorithms like GPTQ

- Only 1 H100 needed for running GPT-3 with int4 weights
- Paving the way for faster inference with integer operations



WHAT I DID

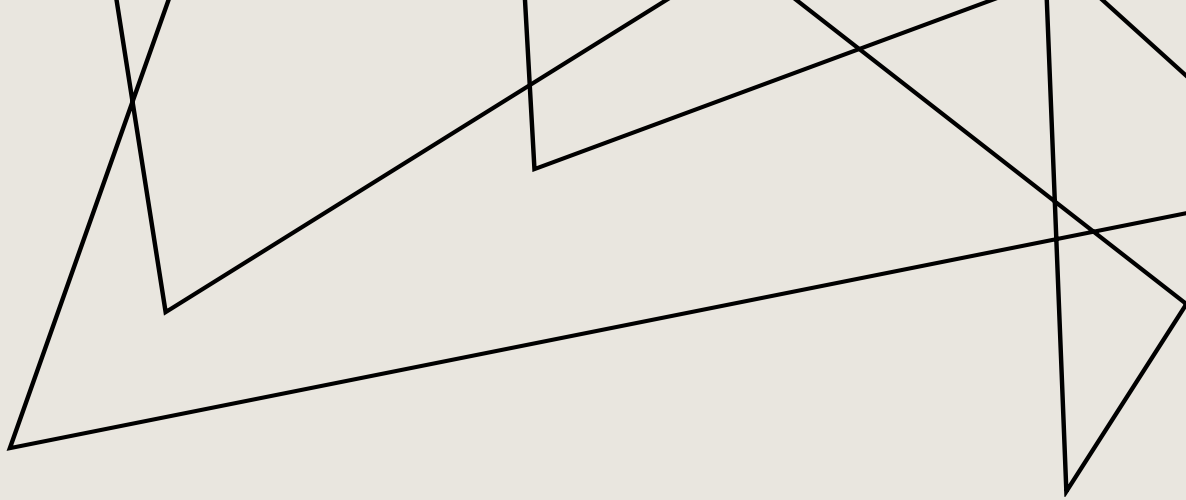
- Re-implemented a 2-3 M parameter transformer
- Trained it on a 10 MB text dataset

```
float32 → int8
```

WHAT I DID

- Re-implemented a 2-3 M parameter transformer
- Trained it on a 10 MB text dataset
- Implemented two quantization methods for int8
- Quantization it using naive rounding
- Quantization using GPTQ

```
float32 → int8
```



WHAT I DID

- Re-implemented a 2-3 M parameter transformer
- Trained it on a 10 MB text dataset
- Implemented two quantization methods for int8
- Quantization it using naive rounding
- Quantization using GPTQ
- Comparison, evaluation, demo, ...

`float32 → int8`

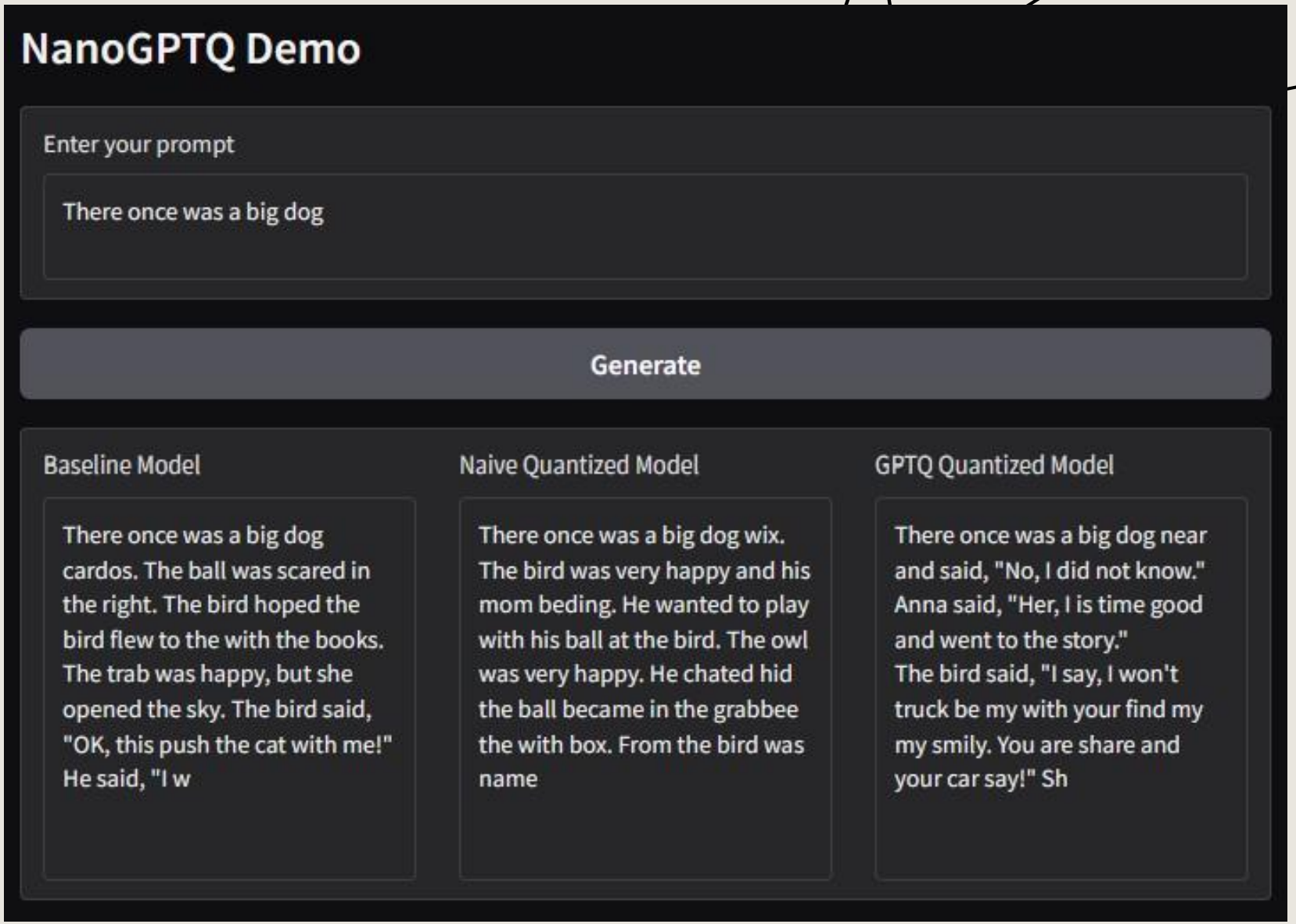
RESULTS

Model Variant	Perplexity (lower is better)	
Baseline (float32)	2.76	
Naive Quantization (int8)	2.78	
GPTQ (int8)	2.76	

RESULTS

Model Variant	Perplexity (lower is better)	Model size
Baseline (float32)	2.76	31.9 MB
Naive Quantization (int8)	2.78	2.9 MB
GPTQ (int8)	2.76	2.9 MB

DEMONSTRATION



INSIGHTS

- Re-implementing NanoGPT was difficult but insightful
- Good exercise for anyone wanting to understand autoregressive transformers

INSIGHTS

- Re-implementing NanoGPT was difficult but insightful
- Good exercise for anyone wanting to understand autoregressive transformers

- Training a 2-3 M parameter model on a tiny notebook GPU was also difficult but insightful
- Learned how to manage and balance expensive training runs

NanoGPTQ

Philipp Sepin