Large Language Models

from first principles to SOTA

A short review

Xavier Martinet Meta

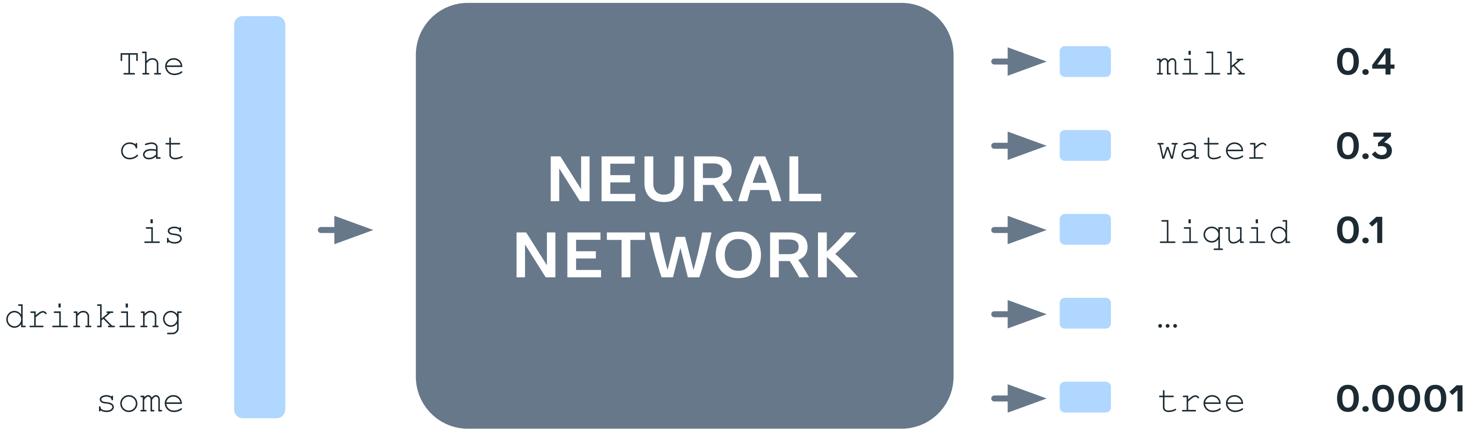
Research Engineer

5 years at Meta
(FAIR then GenAI)

Author of LLaMA, Llama 2,
Llama 3, Llama 3.1

What is a (Large) Language Model?

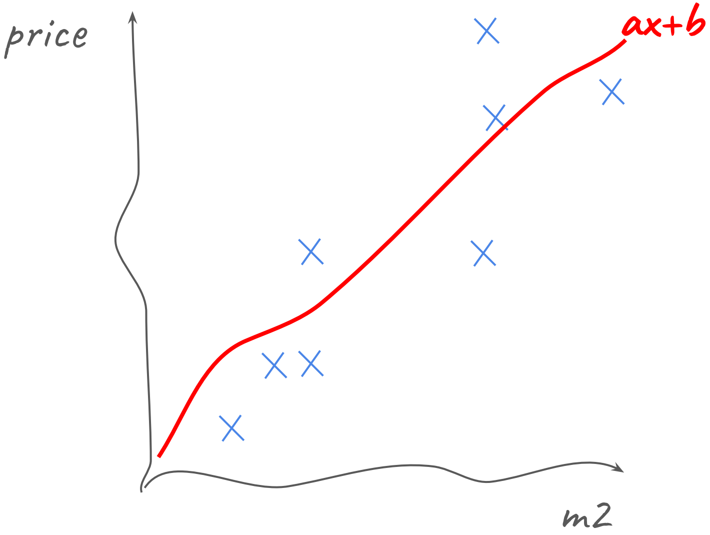
Just a function that outputs the probability over all possible "words"*



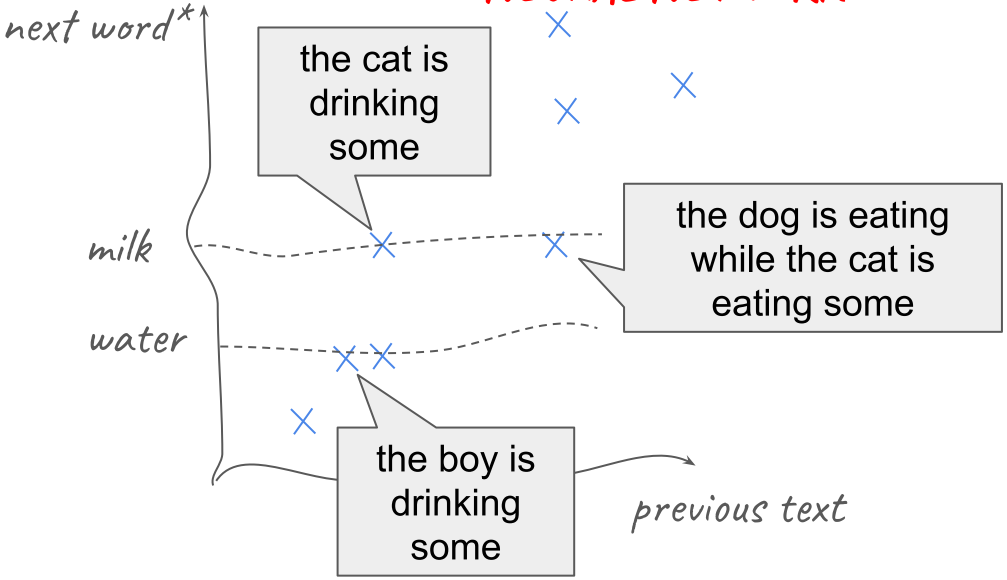
*"tokens", not "words", to be more specific

How does it work?

It fits the parameter of a function to a set of data



The function that fits the data is the **NEURAL NETWORK**



the cat is drinking some

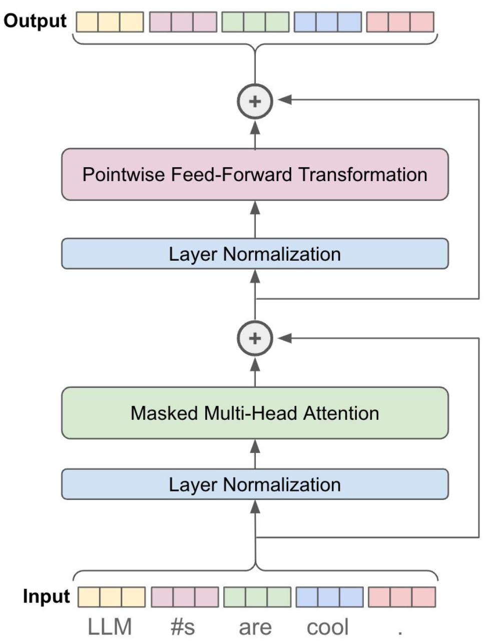
the dog is eating while the cat is eating some

the boy is drinking some

* "tokens", not "words", to be more specific

Nowadays they are based on the *Transformer* architecture

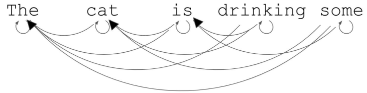
Especially its autoregressive "decoder-only" flavor


Output
Pointwise Feed-Forward Transformation
Layer Normalization
Masked Multi-Head Attention
Layer Normalization
Input
LLM
#s
are
cool
.

Other tricks:

Causal Masking

Tokens can only attend "the past"



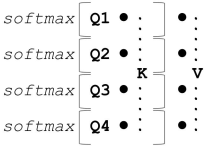
Rotary Positional Encoding

Queries and keys are rotated to encode absolute and relative positions

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

Group Query Attention

Several Query Vectors per Key-Value Pairs



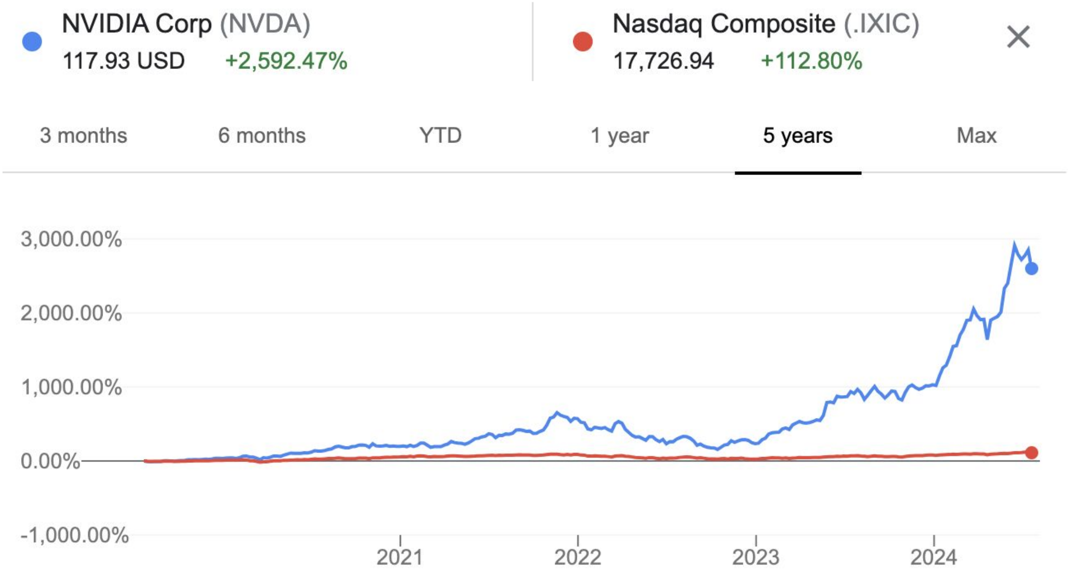
Key-Value Caching (inference)

K-V tensors are not recomputed at decoding time

Credit: https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse

Fitting the function is called Training.
Applying it is called Inference

And it needs GPUs. Like a lot.



● NVIDIA Corp (NVDA)
117.93 USD +2,592.47%

● Nasdaq Composite (.IXIC)
17,726.94 +112.80%

LLMs are so large that memory is scarce on GPU

Llama 2 7B

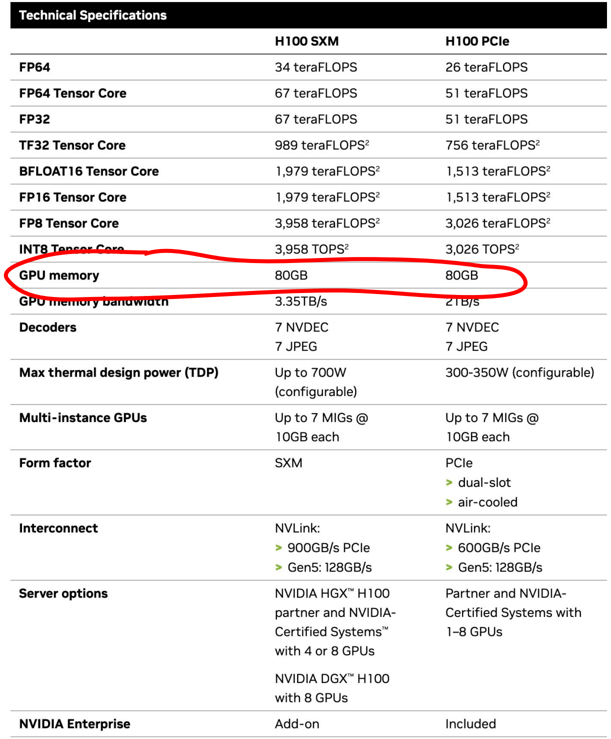
- if weights are stored as fp32 = 4 bytes per parameter

so 7B*4 = 28GB model size

to be added

- the optimizer state: another 28GB + 28GB
- gradients: another 28GB
- activations: depends on the batch size, but > 28GB

Total > 130GB



Technical Specifications

	H100 SXM	H100 PCIe
FP64	34 teraFLOPS	26 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS
FP32	67 teraFLOPS	51 teraFLOPS
TF32 Tensor Core	989 teraFLOPS[2]	756 teraFLOPS[2]
BFLOAT16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]
FP16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]
FP8 Tensor Core	3,958 teraFLOPS[2]	3,026 teraFLOPS[2]
INT8 Tensor Core	3,958 TOPS[2]	3,026 TOPS[2]
GPU memory	80GB	80GB
GPU memory bandwidth	3.35TB/s	2TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)
Multi-instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 10GB each
Form factor	SXM	PCIe > dual-slot > air-cooled
Interconnect	NVLink: > 900GB/s PCIe > Gen5: 128GB/s	NVLink: > 600GB/s PCIe > Gen5: 128GB/s
Server options	NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1–8 GPUs
NVIDIA Enterprise	Add-on	Included

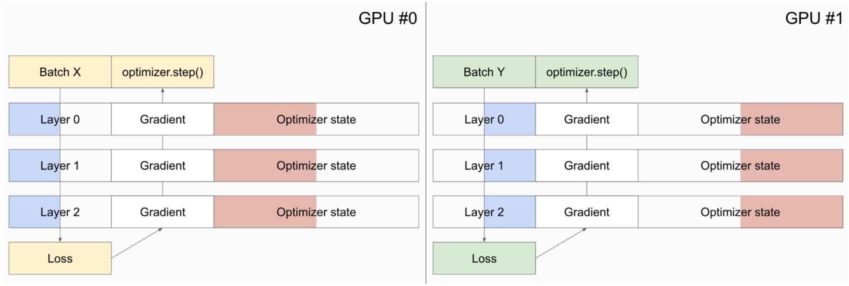
First trick: Activation Checkpointing

Activations are recomputed during the backward pass (trading memory against computation)

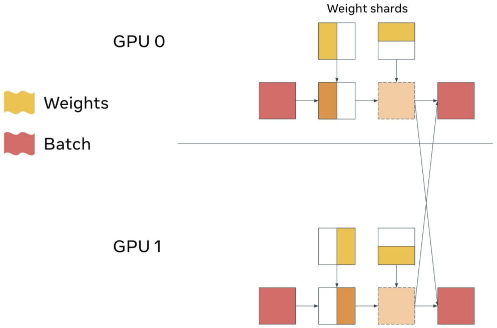
Sharding is necessary to have an LLM fits into the hardware

Memory footprint is traded against networking workload

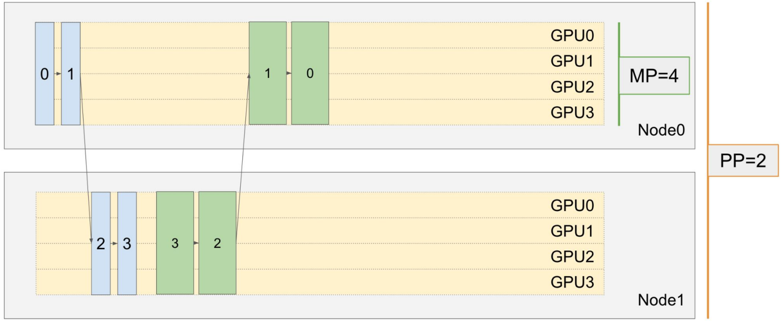
Fully-Sharded Data Parallelism



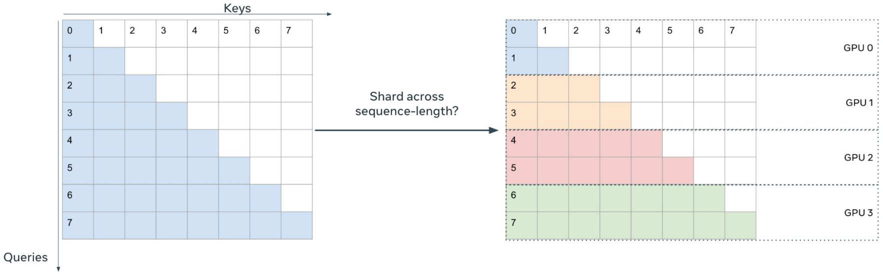
Tensor Parallelism



Pipeline Parallelism



Context Parallelism

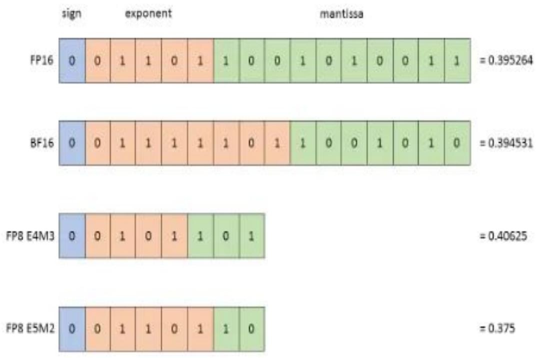


Training in Fp8

Because 8 fingers is all you need

Fp8 is good for you

- Reduces memory footprint of the model
- Augments training speed (up to 30%)



	sign	exponent	mantissa	
FP16	0	0 1 1 0 1	1 0 0 1 0 1 0 0 1 1	= 0.395264
BF16	0	0 1 1 1 1 1 0 1	1 0 0 1 0 1 0	= 0.394531
FP8 E4M3	0	0 1 0 1	1 0 1	= 0.40625
FP8 E5M2	0	0 1 1 0 1	1 0	= 0.375

Necessitates specialized hardware H100 fp8 tensor cores

Numerical Instability:

- gradients over and under flowing
- training divergence

Multi Mixed Precision:

- bf16, fp8, "bf8" (E5M2)
- dynamical scaling during backward pass

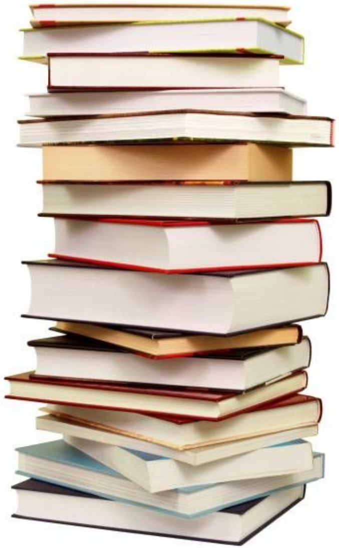
Scaling Techniques:

- JIT scaling / Delayed scaling
- Tensor dimension scaling

Picking the right vocab size and tokenizer training corpus is a big deal for multi-linguality

The Byte-Pair Encoding algorithm is at the basis of modern tokenizers

BPE training corpus



The BPE algorithm clusters characters in sequences according to how frequently they appear in the corpus → Multilingual corpus: frequencies are language dependent and so are clusters obtained from BPE

It stops when the total number of groups is reached: the vocabulary size → Larger vocabulary size: the long tail of less frequent sequences can be tokenized

These groups are called "tokens" → Fewer tokens to encode the same message: less probability for the LLM to go astray

With some finetuning, we can make it chat

Neural Network are good at picking patterns.

Pretraining stage:
predicts the next word

```
I have a dream that one day on the
red hills of Georgia, the sons of
former slaves and the sons of
former slave owners will be able
to sit down together at the table
of brotherhood.
I have  ???
```

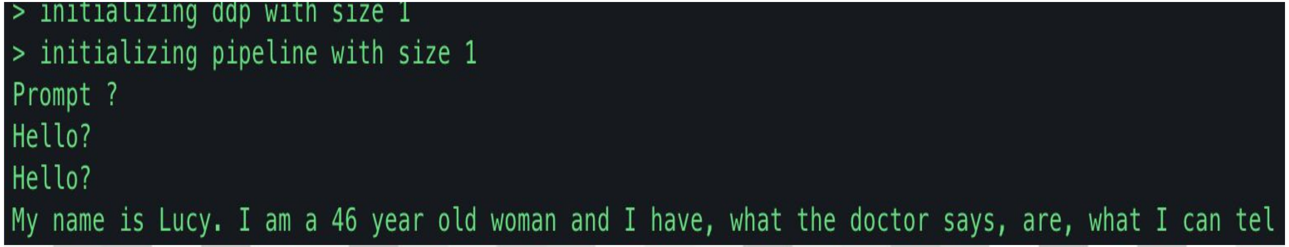
The LLM will keep talking
forever as in a monolog

Finetuning stage:
still predicts the next word

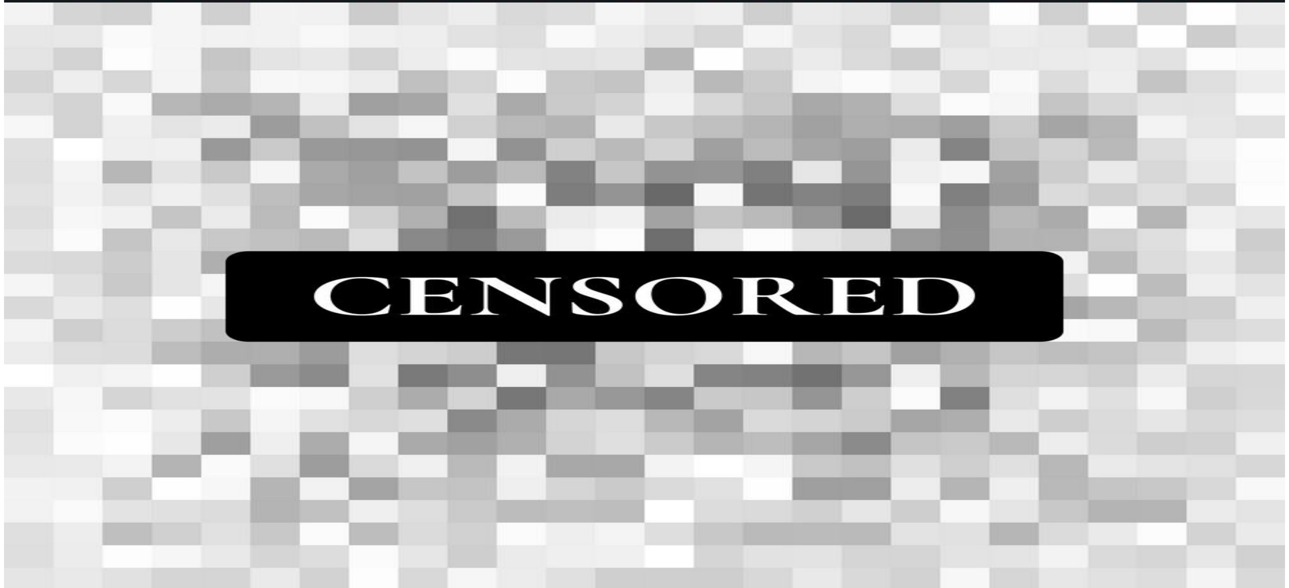
```
<user>hello</user>
<assistant>how can I help you?</assistant>
<user>well, I am looking for a birthday present</user>
<assistant> ???
```

Still the next word to be predicted,
but conditioned on a specific "chat" pattern

But the content from the web, oh boy...

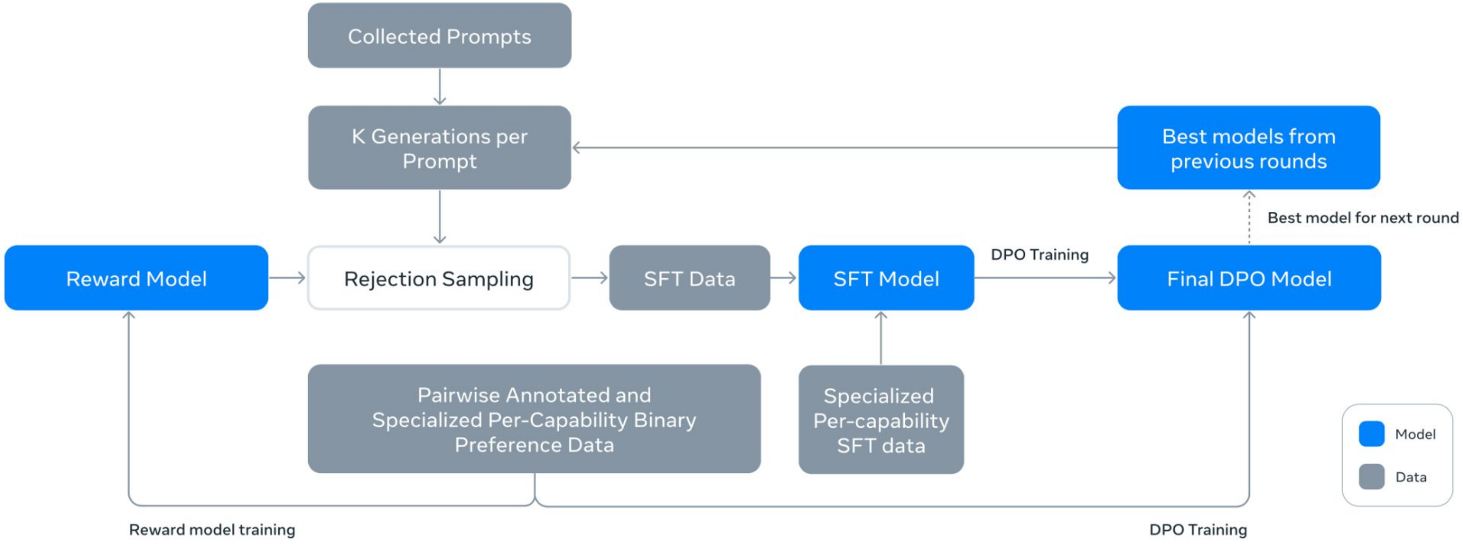


```
> initializing ddp with size 1
> initializing pipeline with size 1
Prompt ?
Hello?
Hello?
My name is Lucy. I am a 46 year old woman and I have, what the doctor says, are, what I can tel
```



RLHF* FTW**

They become better to assess human preferences than humans themselves


Collected Prompts
K Generations per Prompt
Best models from previous rounds
Best model for next round
Reward Model
Rejection Sampling
SFT Data
SFT Model
DPO Training
Final DPO Model
Pairwise Annotated and Specialized Per-Capability Binary Preference Data
Specialized Per-capability SFT data
Model
Data
Reward model training
DPO Training

Credit:
The Llama 3 Herd of Models, Llama Team, AI @ Met

**Reinforcement Learning from Human Feedback*   ***For The Win*

Chat-tuned models behave as expected

HHH: Helpful, Honest, Harmless

PROMPT: What to visit in New York City?

Pretrained model completion

Travel Forums North America What to visit in New York City?

1. Posted by michelle123 (Budding Member 3 posts) 11y Star this if you like it!

I'm going to New York City for the first time this weekend.

Chat-tuned model completion

New York City is home to many iconic landmarks and attractions, such as the Statue of Liberty, Central Park, Times Square, Empire State Building, and more. The city also has a vibrant arts scene with numerous museums and galleries, including the Metropolitan Museum of Art and the Guggenheim Museum. If you're interested in trying new foods, NYC offers diverse dining options from around the world, ranging from street vendors to Michelin-starred restaurants. You could also catch a Broadway show or visit one of the many neighborhoods that offer unique experiences like Little Italy, Chinatown, and Williamsburg.

Llama through time

	LLaMA Feb. 2023	Llama 2 July 2023	Llama 3 April 2024	Llama 3.1 July 2024
Sizes	7B, 13B, 32B, 65B	7B, 13B, 70B	8B, 70B	8B, 70B, 405B
# tokens	1.4T	2T	15T	15T
Context length	2k	4k	8k	128k
Changelog	● Pretrained only ● 32k vocab size ● RoPE ● RMS pre attention	● Chat-tuned ● GQA ● RLHF loop w/ PPO ● Built-in safety	● RLHF loop w/ DPO ● 128k vocab size	● 8 languages ● Focus on code and reasoning ● Tool capabilities

LLMs are trained to be generalist assistants, in-context learning makes them suitable for specific needs

Expensive finetuning is not necessary for money-tight use cases

Fewshot learning

LLM are pattern-catching experts: they can grab what is expected from them if they are shown examples

Retrieval Augmented Generation (RAG)

Potentially relevant chunks of text are extracted from a database, and fed in-context

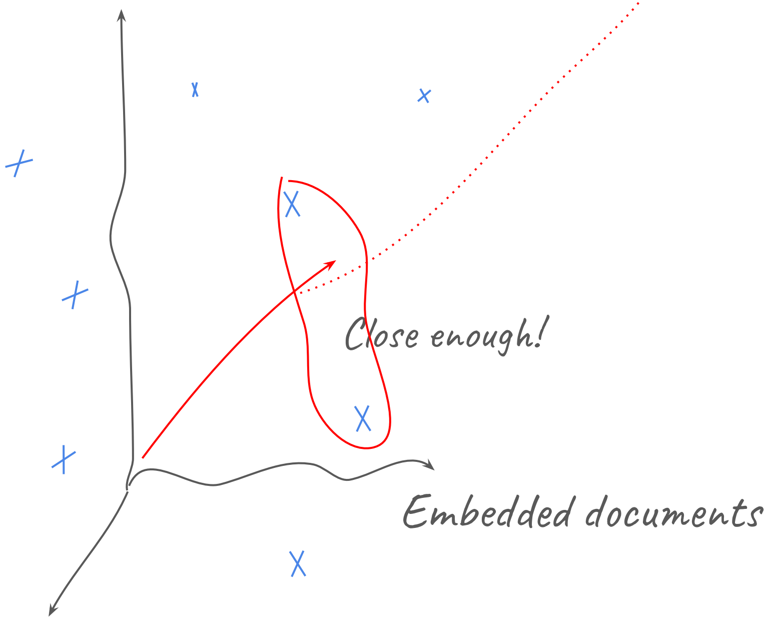
Long context finetuning

Frequencies in RoPE are increased and longer training samples used to increase the maximum context length

RAG

Because LLMs too want to use Google

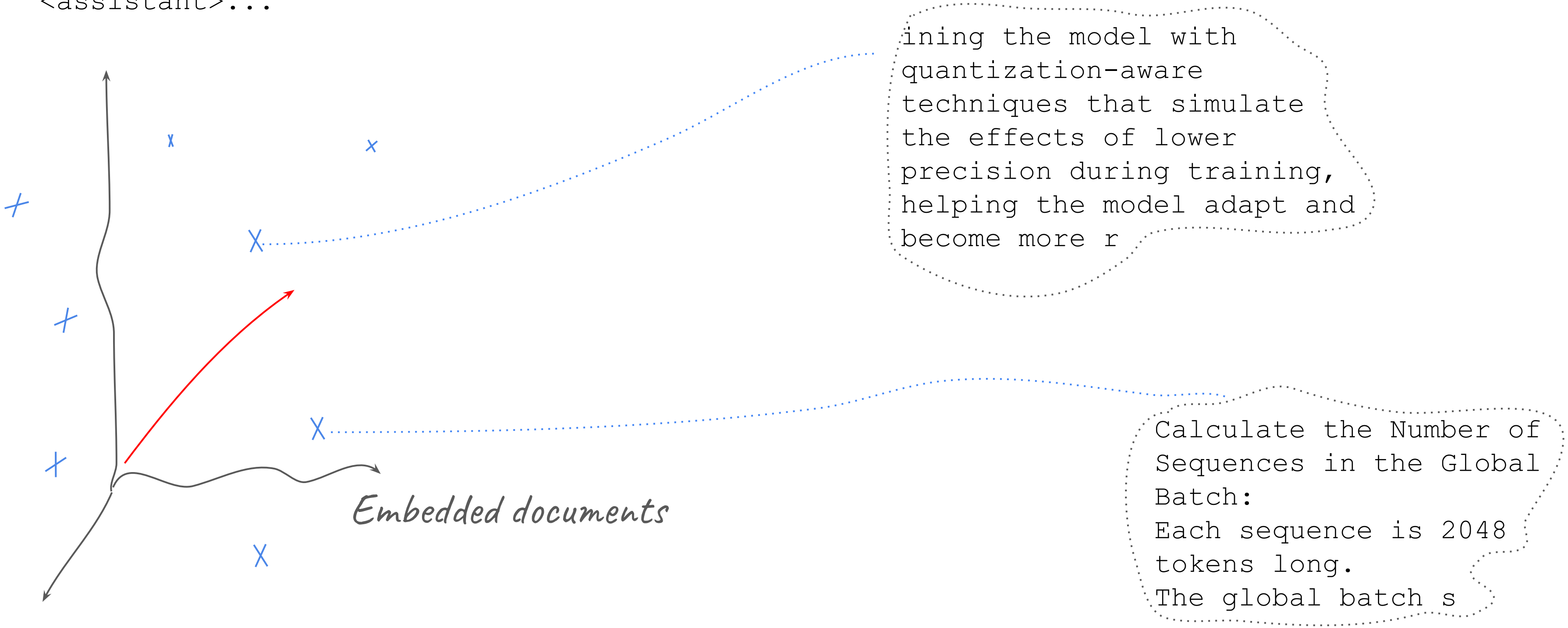
```
<user>Can you tell me how to train an AGI model, with sources?
</user>
<assistant>...
```



RAG

Because LLMs too want to use Google

```
<user>Can you tell me how to train an AGI model, with sources?
</user>
<assistant>...
```



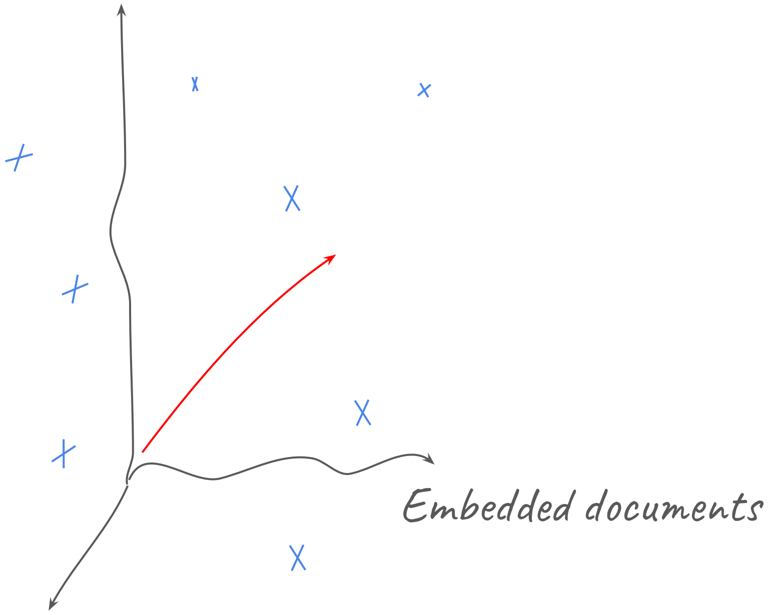
ining the model with quantization-aware techniques that simulate the effects of lower precision during training, helping the model adapt and become more r

Calculate the Number of Sequences in the Global Batch:
Each sequence is 2048 tokens long.
The global batch s

RAG

Because LLMs too want to use Google

```
<user>Can you tell me how to train an AGI model, with sources?
</user>
<assistant>...
```

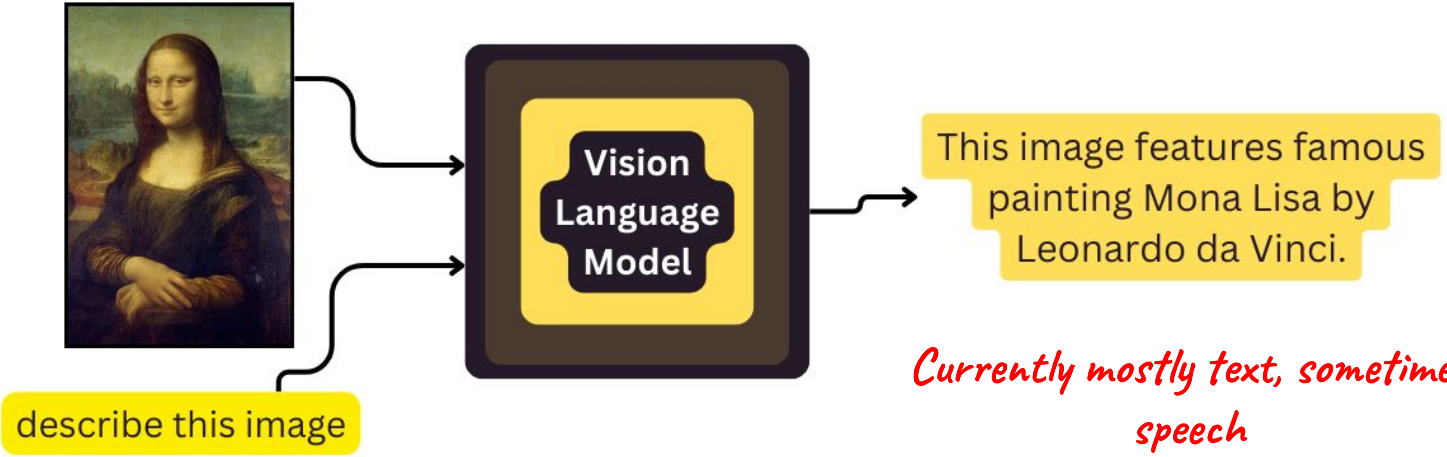


Retrieval of document procedure

- *Soft retrieval*
- *Cross attention mechanism*
- *Reranking of the results*
- *Overloading of the prompt*
- *Refines the model generation*
- *Provides sources*

The revolution might not be televised*, but LLMs will definitely see

The faculty of having non-textual inputs and outputs is called *multimodality*



describe this image

Vision Language Model

This image features famous painting Mona Lisa by Leonardo da Vinci.

Also speech, videos

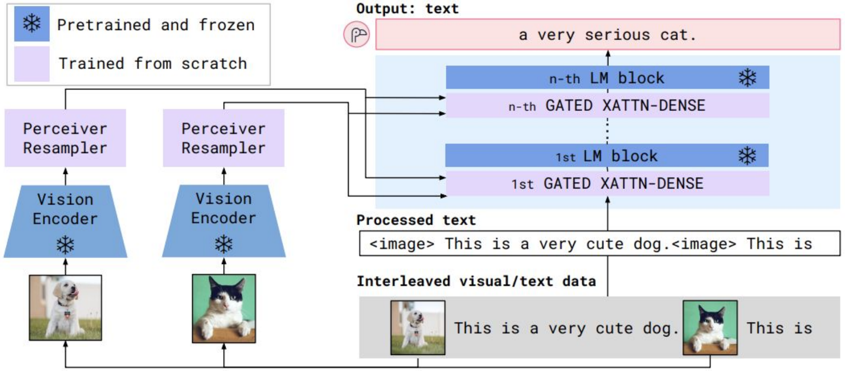
Currently mostly text, sometimes speech

But no images or videos :(

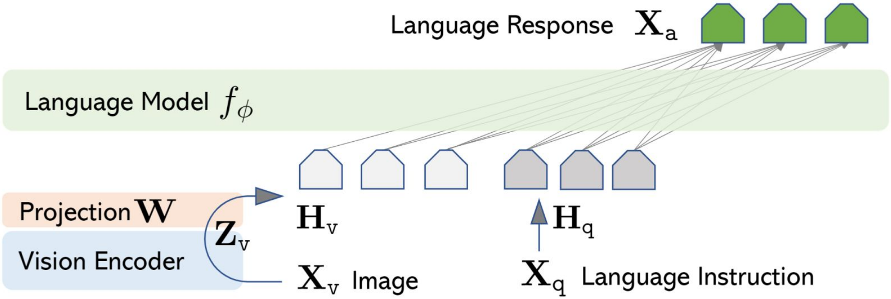
**The Revolution Will Not Be Televised, song by Gil Scott-Heron, 1970*

Two architectural paradigms drive this trend: Cross-Attention and Early Fusion

Cross Attention
"Flamingo-like"



Early Fusion
"LLaVA-like"



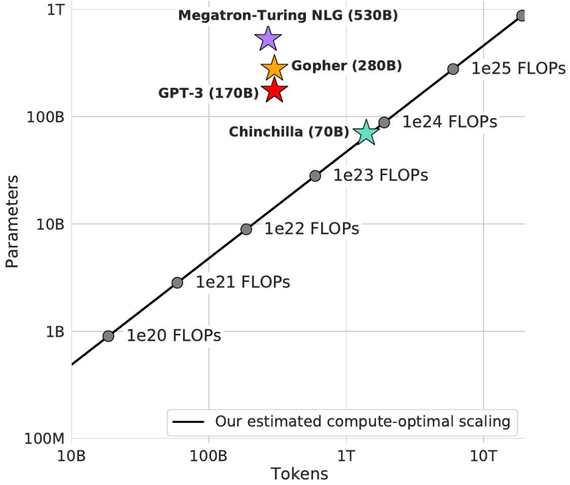
Credit:
Flamingo: a Visual Language Model for Few-Shot Learning, DeepMind

Credit:
https://llava-vl.github.io/

What's next in the LLM world?

More generalists or more specialized, but with less prompt engineering

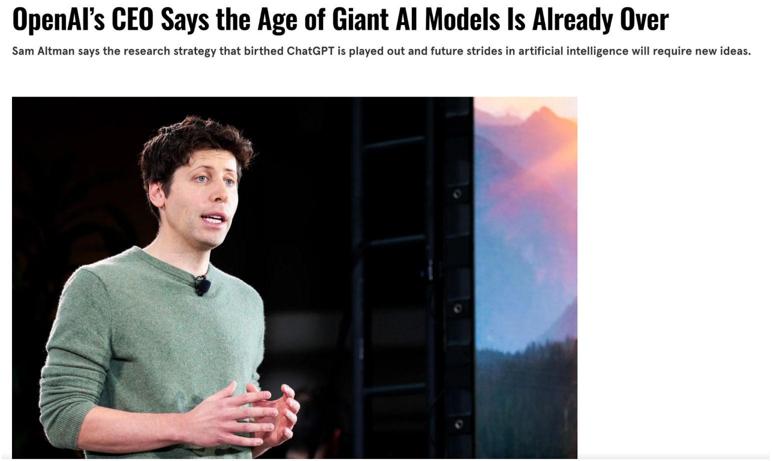
Larger is better



but diminishing returns?

or

Small is beautiful



OpenAI's CEO Says the Age of Giant AI Models Is Already Over

Sam Altman says the research strategy that birthed ChatGPT is played out and future strides in artificial intelligence will require new ideas.

but falling short of AI promises?

MoE: Model Specialization

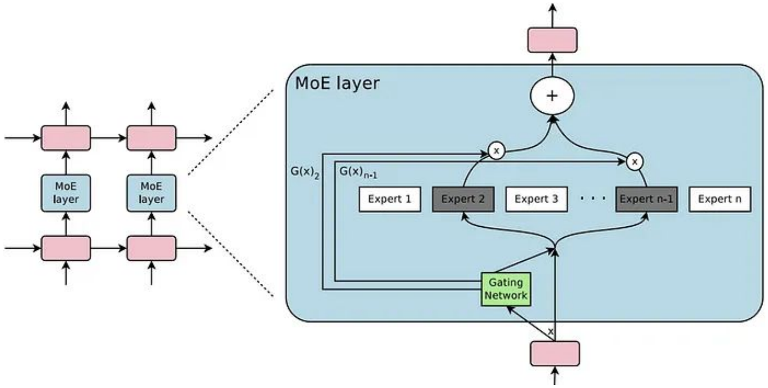
Reduce model footprint and/or improve its performance

Many Flavors of routing scheme

Sequence/Token Routing

Expert/Token Choice

Memory/performance trade-off



Sparse MoE/Dense MoE

$$\frac{\partial \mathcal{L}}{\partial W_r} := \nabla_0 + \nabla_1, \text{where } \nabla_0 = \sum_{\boldsymbol{I}_i} g(\boldsymbol{\pi}_{\boldsymbol{I}_i} f_{\boldsymbol{I}_i}(\boldsymbol{x})) \frac{\partial \boldsymbol{\pi}_{\boldsymbol{I}_i}}{\partial W_r} \text{ and } \nabla_1 = \sum_{\boldsymbol{I}_i} \boldsymbol{\pi}_{\boldsymbol{I}_i} \frac{\partial g(\boldsymbol{\pi}_{\boldsymbol{I}_i} f_{\boldsymbol{I}_i}(\boldsymbol{x}))}{\partial W_r}$$

Differentiability restoring tricks (STE, Renormalization)

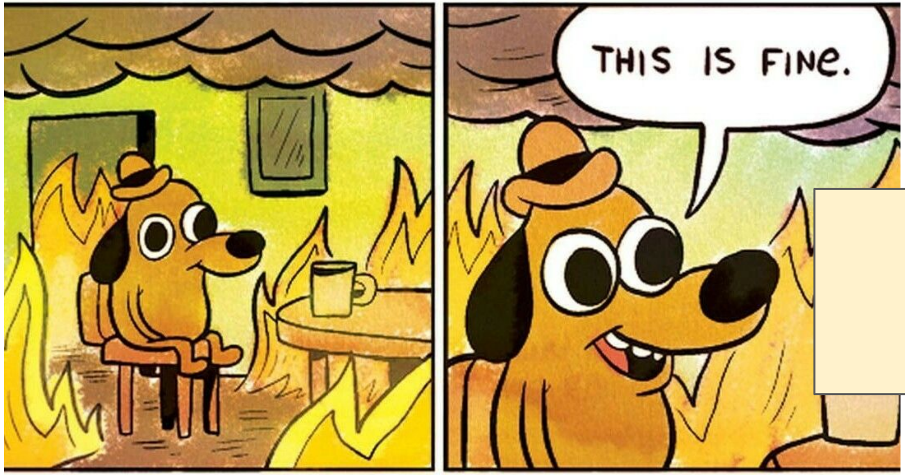
Majority voting, pure expert, specialization

Model merging, ensemble techniques (BTM)

AI should not only to answer questions, but also take actions

Agentic behavior may be the trillion dollar business opportunity that justifies the gigantic investments

PROMPT: Help me navigate to a shirt
that has this on it.

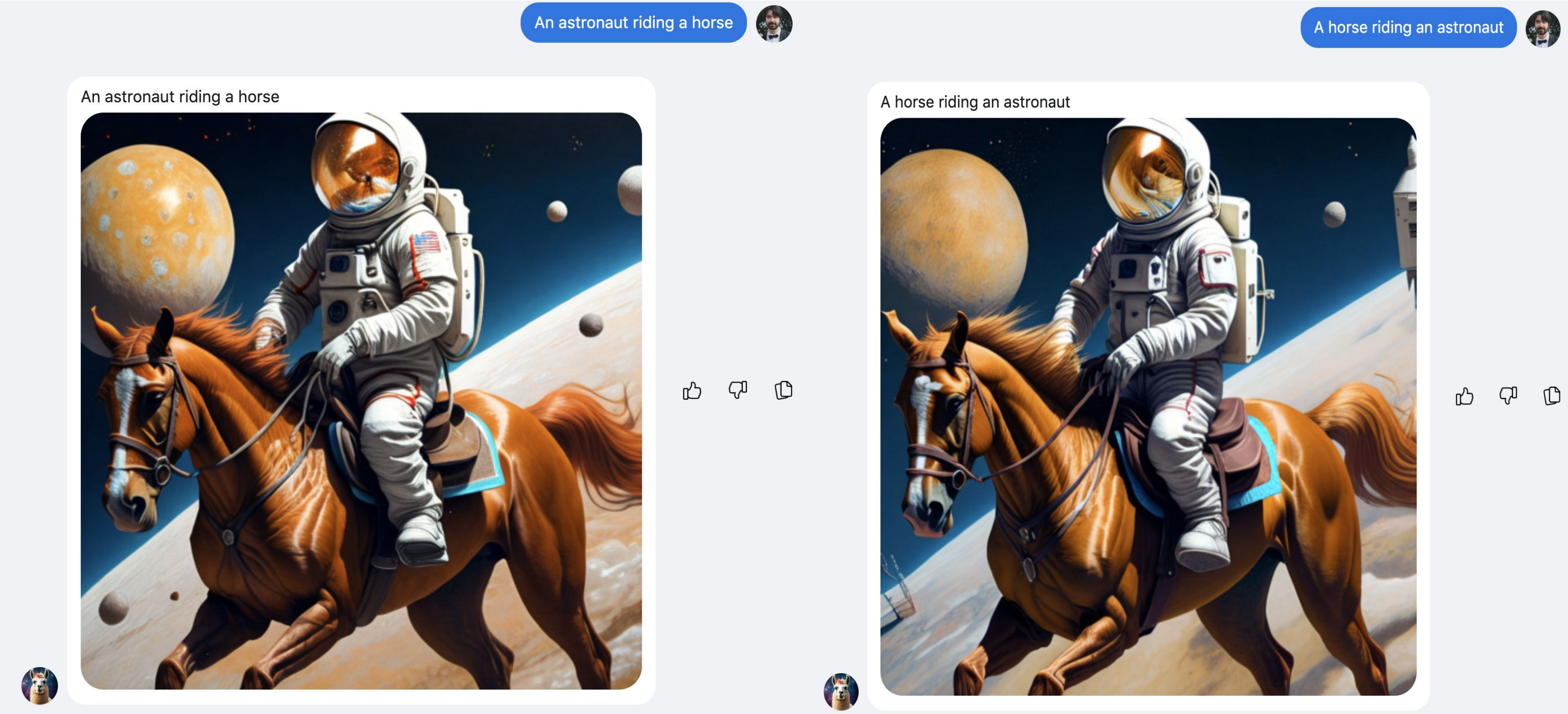


VIDEO

There is still a long way to AGI*

Are Neural Networks smart parrots, or do they possess a real understanding?

Example from a diffusion model, not an LLM, but illustrative of the challenges in the domain



An astronaut riding a horse

An astronaut riding a horse

A horse riding an astronaut

A horse riding an astronaut

*Artificial General Intelligence, loosely defined concept considered as the Graal of AI

The wise man doesn't give the right answers, he poses the right questions

Claude Lévi-Strauss