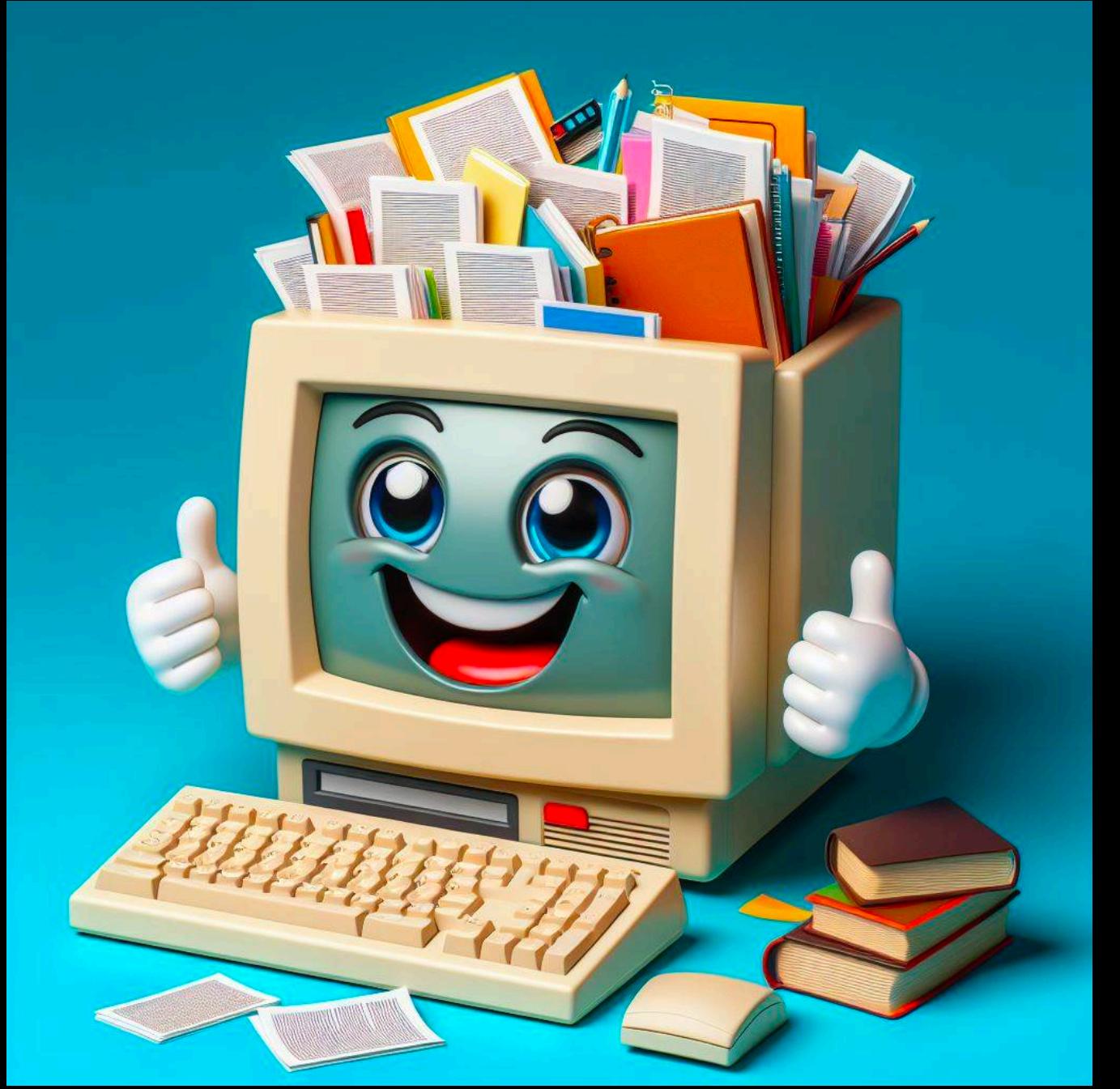


**Days since last AI breakthrough:**  
0

# LLMs. applications, implementations and the latest developments

tollef jørgensen, {today}.



a sarcastic smiling toy computer  
consuming all the worlds documents,  
books, and more

# Applications

- Just use LLMs for everything!
- ... but what is the goal behind **using an LLM for you or your company?**



# Applications

- **Search?** *What did person X say about Y in this news article?*
  - Regex, TF-IDF, dependency trees, named entities, embeddings.
  - Or just google it
- **Ranking and retrieval?** *From a query, what are the best matching documents?*
  - BM-25, statistical models, semantic search with embeddings
- **Clustering?** *What topics are talked about in this document*
  - K-means, Locality-Sensitive Hashing, embeddings + dimensionality reduction.

Matt Shumer   
@mattshumer\_

Gemini 1.5 just found the speaker of a single sentence out of the entire Harry Potter and the Order of the Phoenix book.

Over 360,000 tokens, and it got it right!

GPT-4 can only handle 128,000, and does so poorly.

CK  @cyprianpl · 16h  
Google search engine does this in 0.46 seconds

"look, you don't understand what it was like after it happened"

All Images Videos Perspectives Shopping More ▾

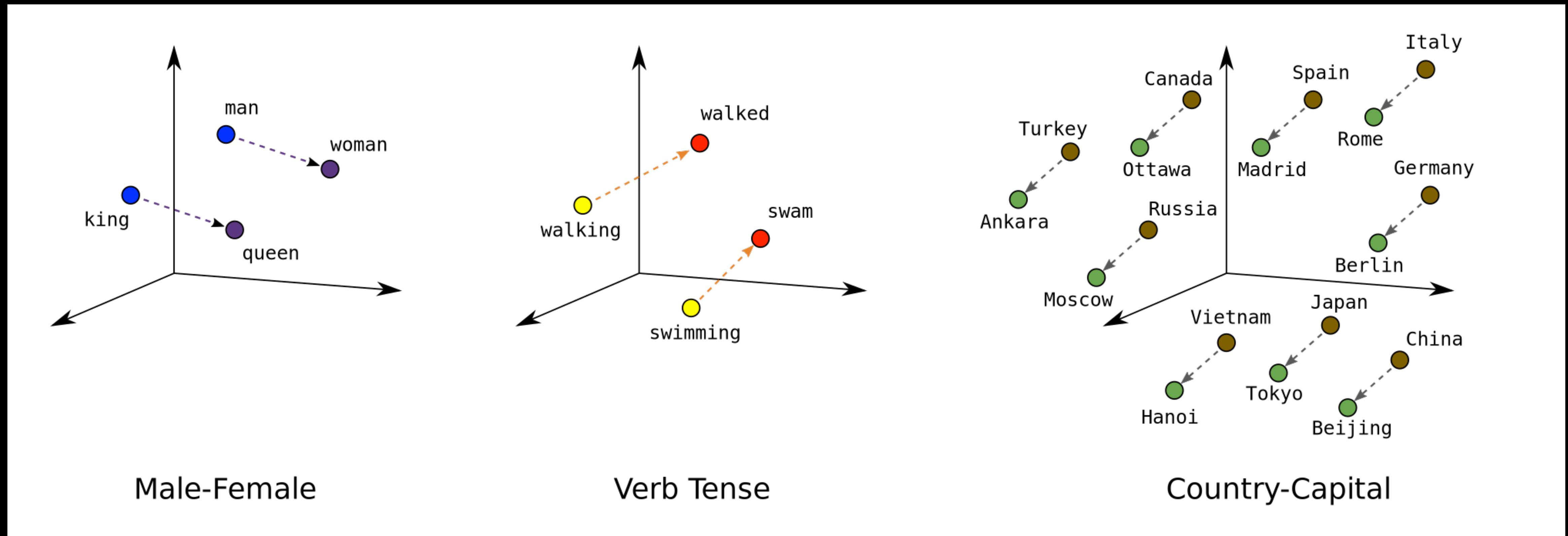
Novel Movie Read

About 2,530 results (0.46 seconds)

"Look, you don't understand what it was like after it happened," said Hermione quietly. "You arrived back in the middle of the lawn clutching

# Embeddings

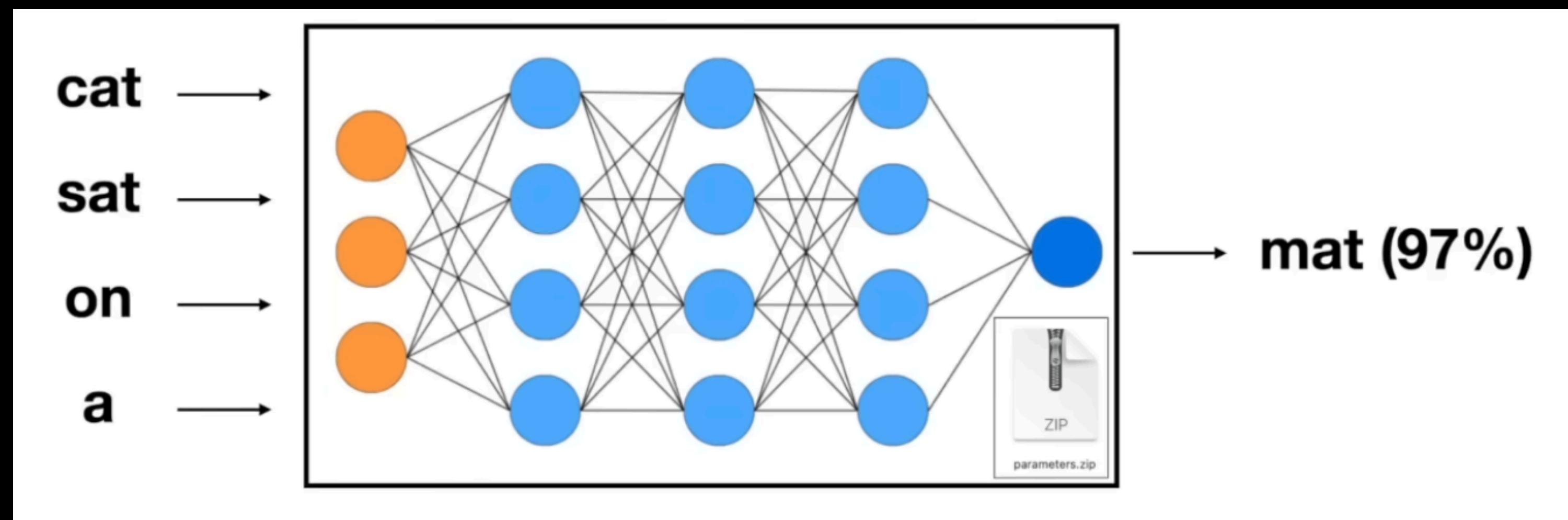
[0.9, 0.3, 0.5] - [0.4, 0.3, 0.6]



Representations that capture the meaning and context of words

# Language models

- ...Many definitions, but let's go with:
  - A model that predicts the likelihood of a token (word) in a given context (sentence)
- *Large* language models?
  - More data! More parameters!



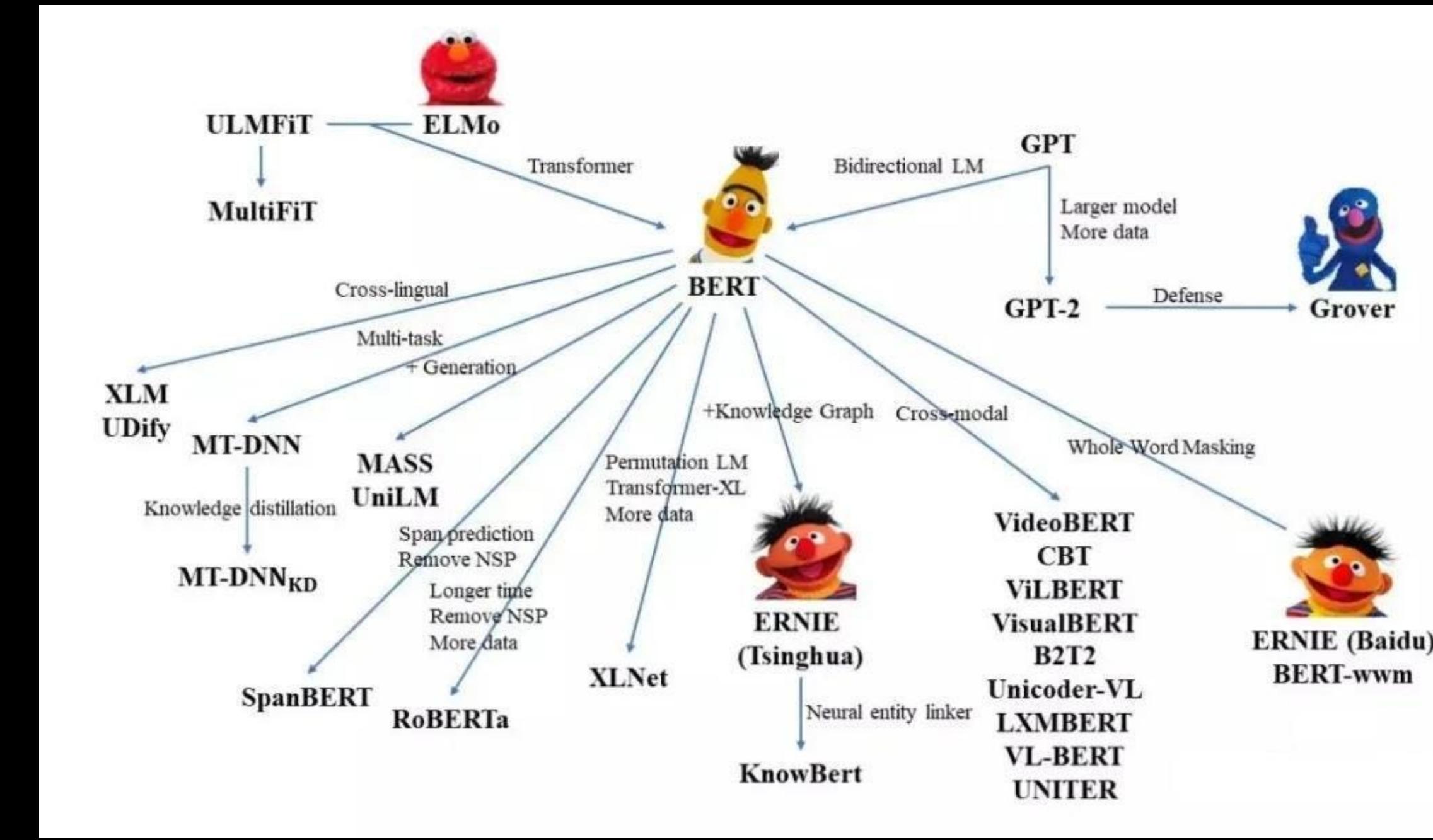
# Misconceptions. An LLM...

- *knows* nothing (unless data is added into its context/prompt)
  - next word prediction wrapped in a fancy dress
- has no idea what words are
  - IDs from a *tokenizer*
    - Created from commonly occurring characters
- has no personality (it's all prompts: "act as a helpful assistant")
  - no, they are not sentient

Tokens	Characters
6	16
	tollef jørgensen
Tokens	Characters
6	26
	has no idea what words are
	[4752, 912, 4623, 1148, 4339, 527]
Tokens	Characters
3	
	2340785

# Language models

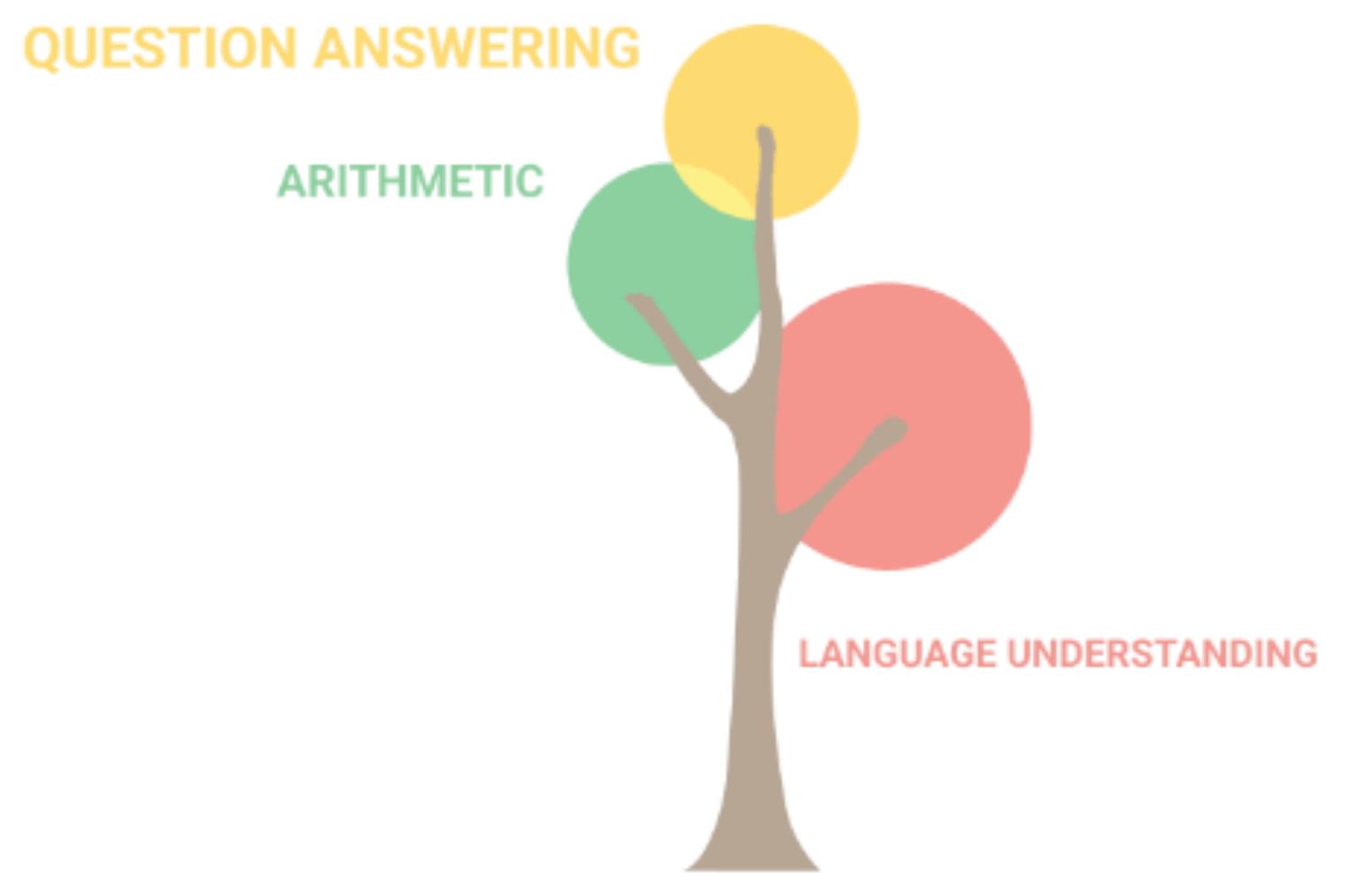
- Modern:
  - Machine learning/statistical models
  - Neural nets: RNNs and LSTMs
- Since 2018:
  - The transformer architecture and related models
  - Google's BERT kickstarted this new phase
    - ... and launched a whole sesame street model-naming campaign



# GPT

- GPT, or *generative pre-trained transformer*, is nothing new
  - The public mostly became aware of the term after the release of ChatGPT in November 2022
- GPT-1: 2018 by OpenAI
  - Limited data sources and size (but it was big for its time!)
- ChatGPT:
  - Based on GPT3 (2020), but it's much more than just a base model
  - "Scaling laws" for transformer-based models
    - More data + more parameters = 
    - Emergent behavior

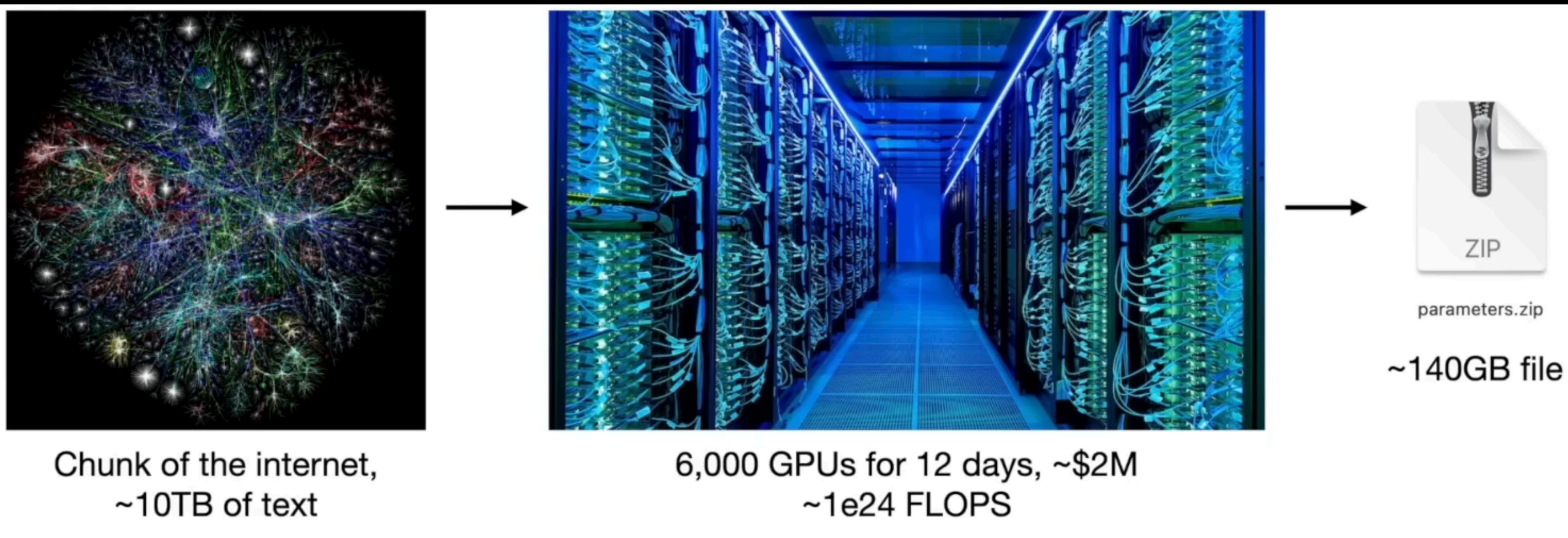
# Language understanding? Reasoning? Emergent behavior...



8 billion parameters

# \$ tiny economical insight \$

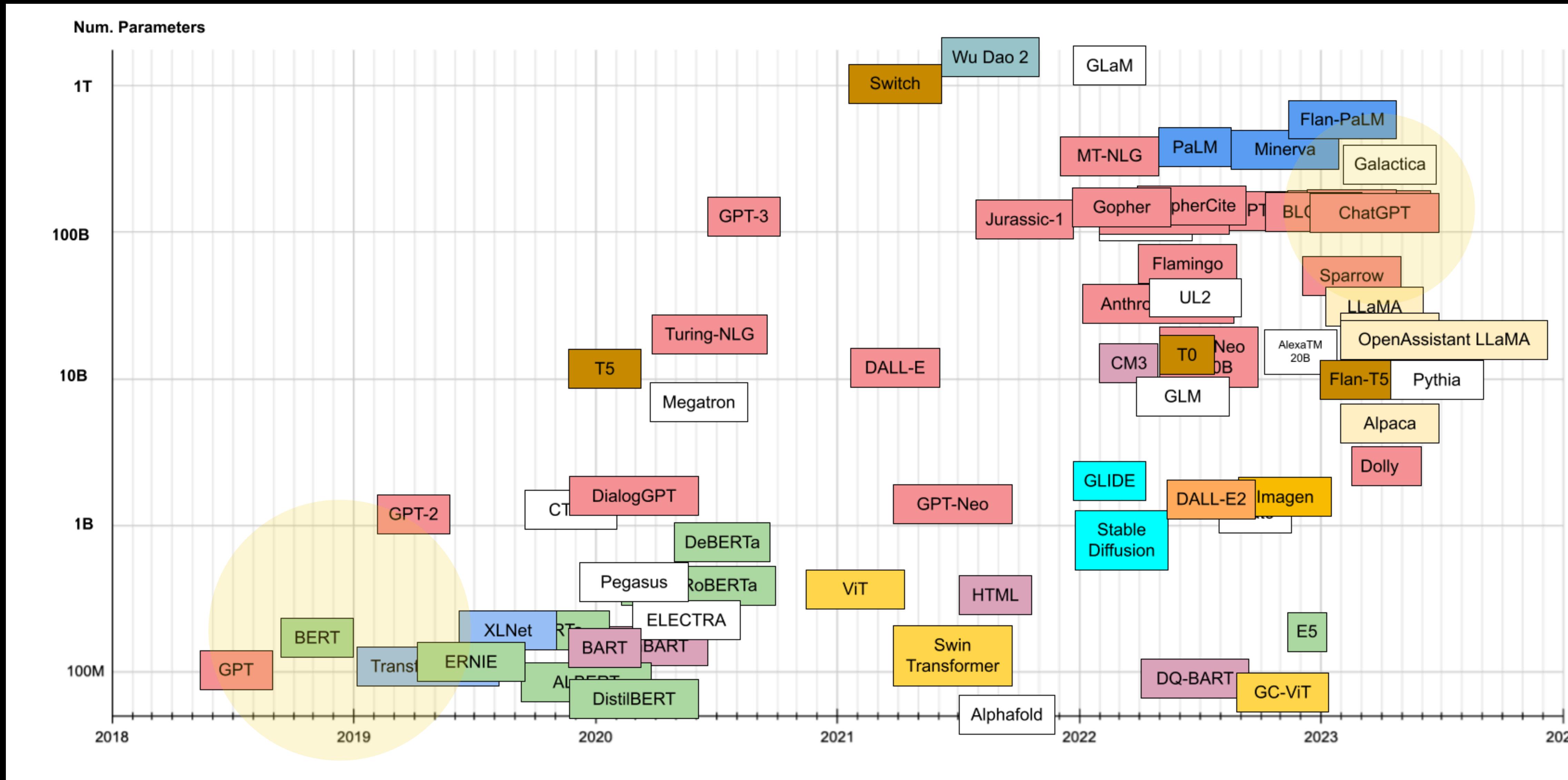
- For reference...
  - Training a 13 billion parameter model on 1.4 trillion tokens (LLaMA)
    - ~\$1 million
  - GPT-4 is known to have cost more than \$100M!



# GPT and friends (or enemies)

- DeepMind
  - Chinchilla family + Gopher, 2021 (gpt-2 with modifications)
    - from 44M to **280B** parameters
- Google
  - PaLM (2022) and PaLM-2 (2023) , 340-540B
  - Bard->Gemini
- OpenAI
  - GPT series (117M, 1.5B, 175B, ???B) - Rumours say up to 1.7 trillion!
- Anthropic
  - Claude models (Claude 3 released March 4, 2024)
- Perplexity
  - Direct ChatGPT competitor with online search (more about this later)

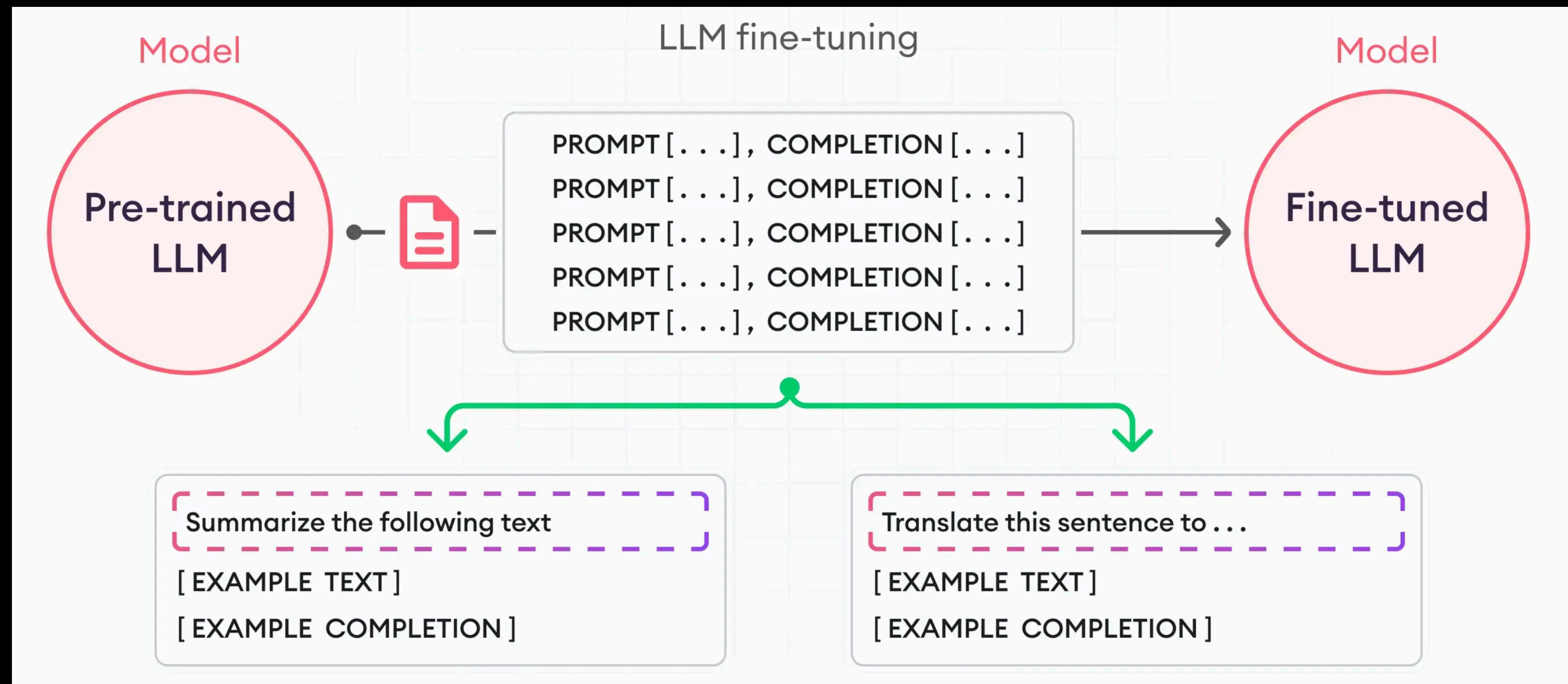
# Timeline, y-axis = N parameters



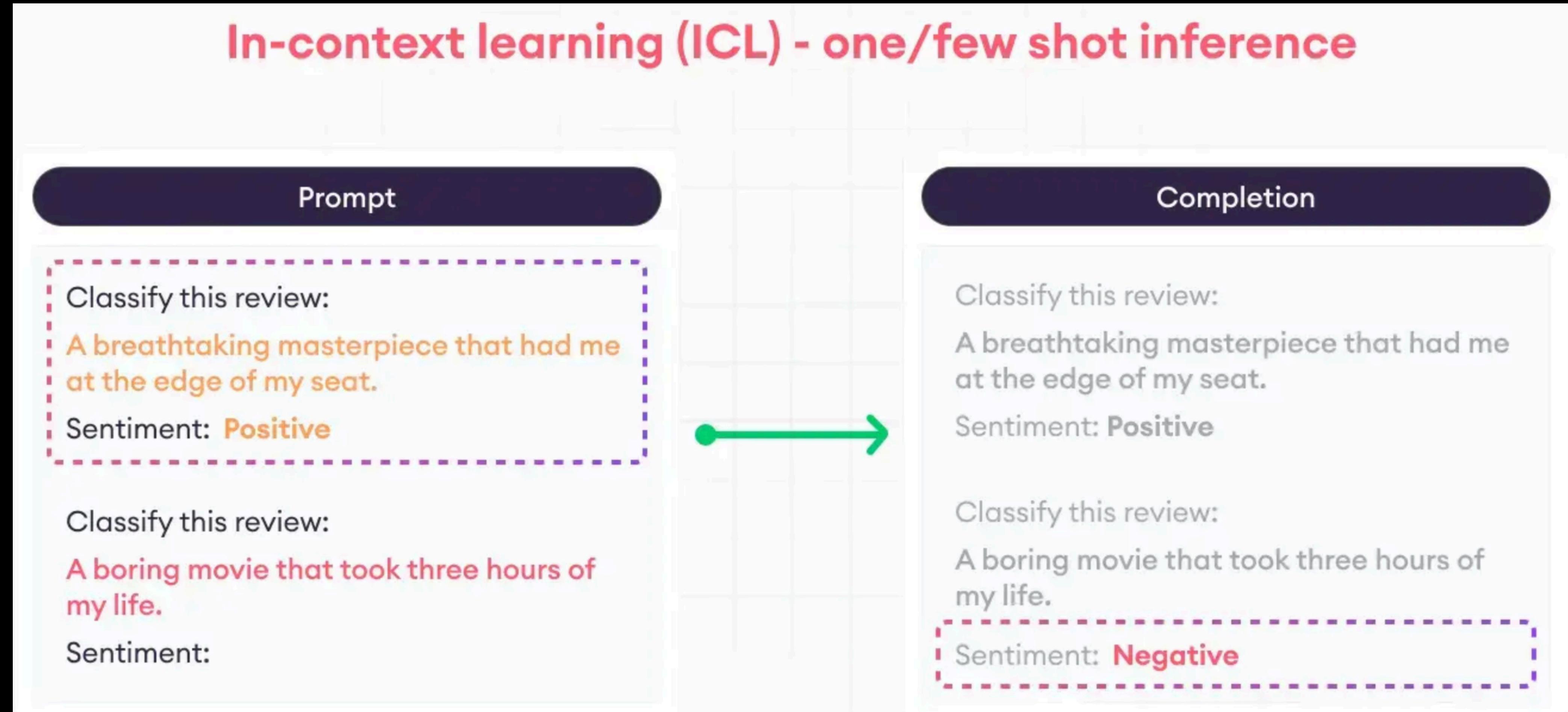
Claude3 Haiku  
turns thousands of  
physical documents  
into structured data

**What is the difference between a base model and what we think of when we hear "LLMs" today?**

# Fine-tuning (often by *instructions*) -> Assistants!



# No fine-tuning: in-context learning



## Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~\$2M, wait ~12 days.
4. Obtain **base model**.

every  
~year

## Stage 2: Finetuning

1. Write labeling instructions
2. Hire people (or use [scale.ai!](#)), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model**.
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

every  
~week

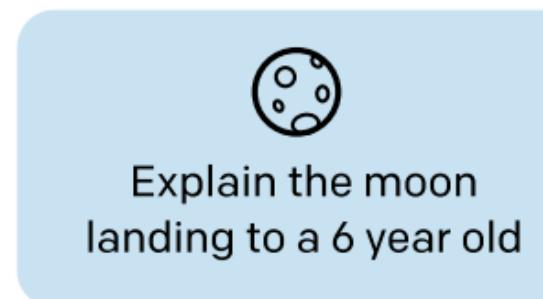
# RLHF and preference tuning

- Reinforcement Learning from Human Feedback
  - This was the secret sauce behind ChatGPT
    - What made it seem like an all-knowing entity

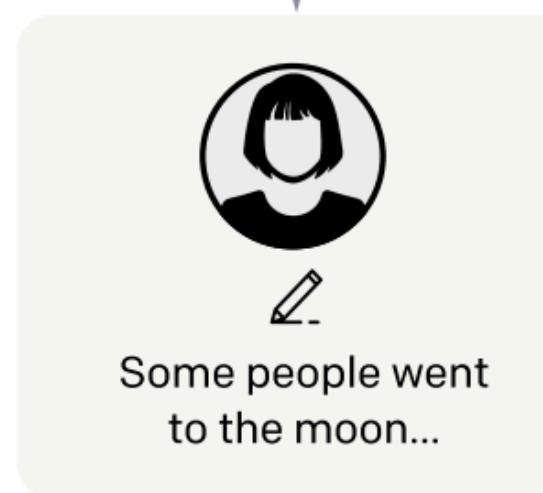
Step 1

**Collect demonstration data, and train a supervised policy.**

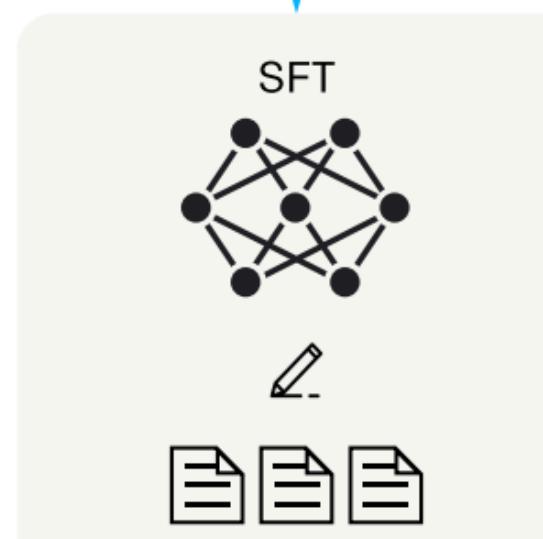
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



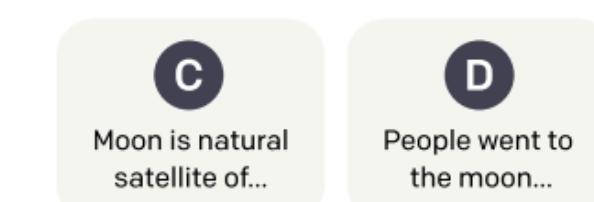
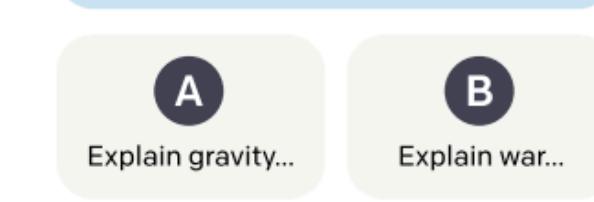
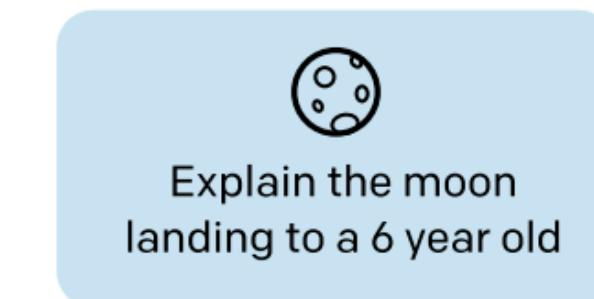
This data is used to fine-tune GPT-3 with supervised learning.



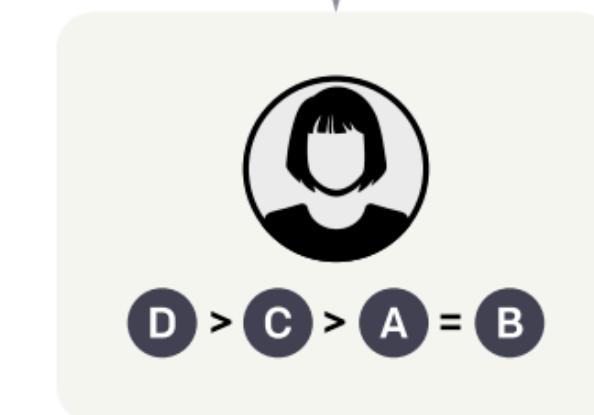
Step 2

**Collect comparison data, and train a reward model.**

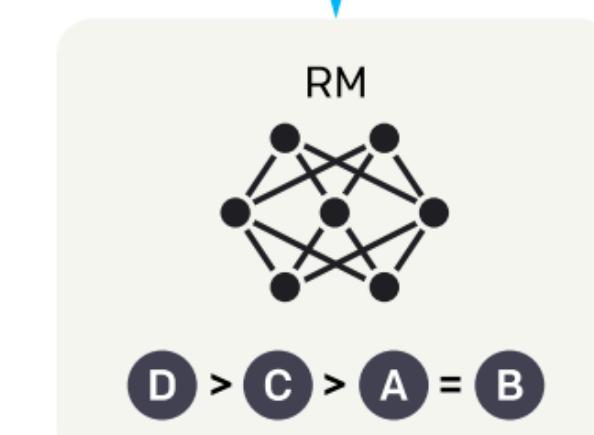
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



D > C > A = B

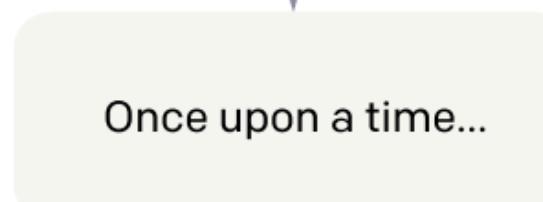
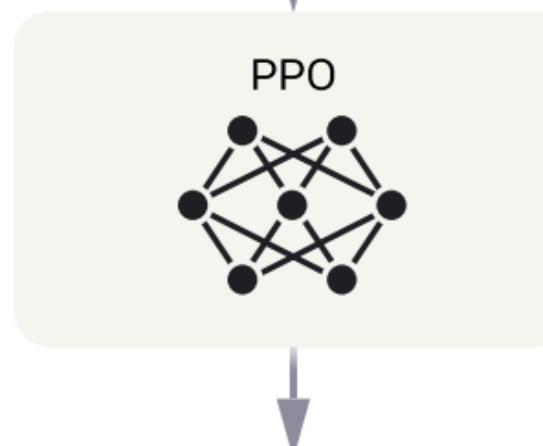
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

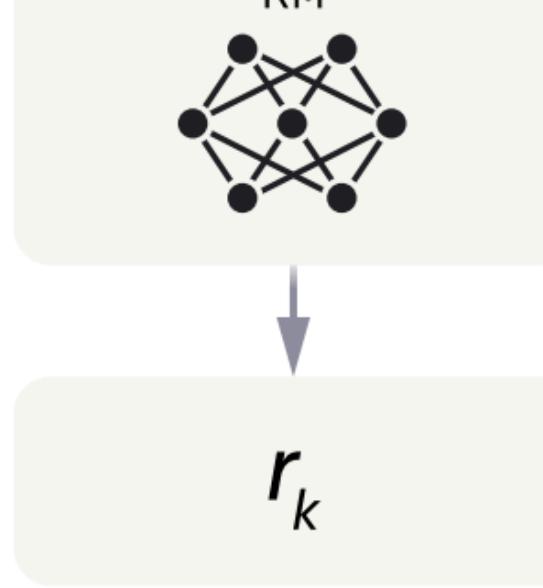
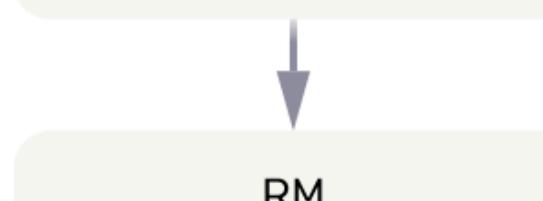
A new prompt is sampled from the dataset.



The policy generates an output.



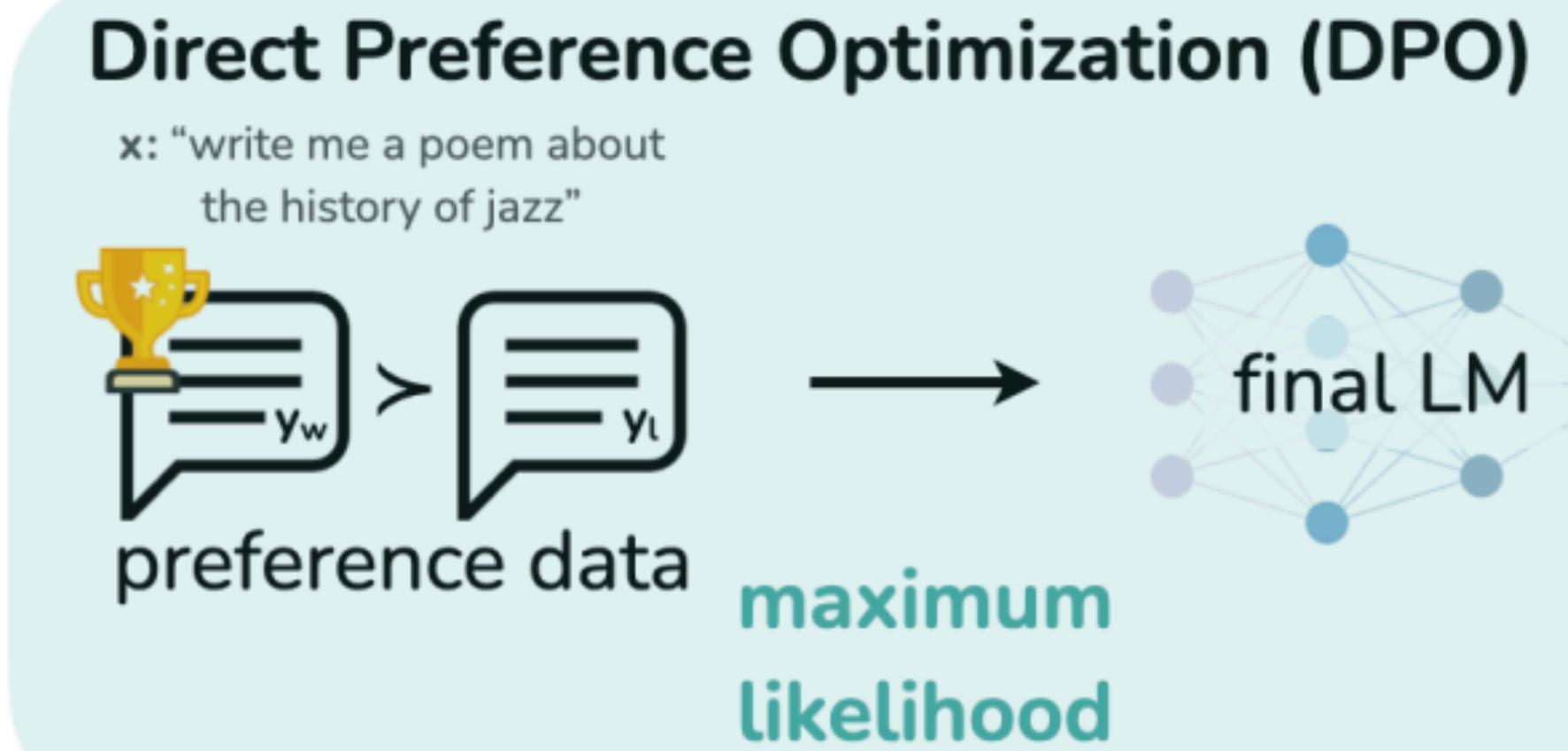
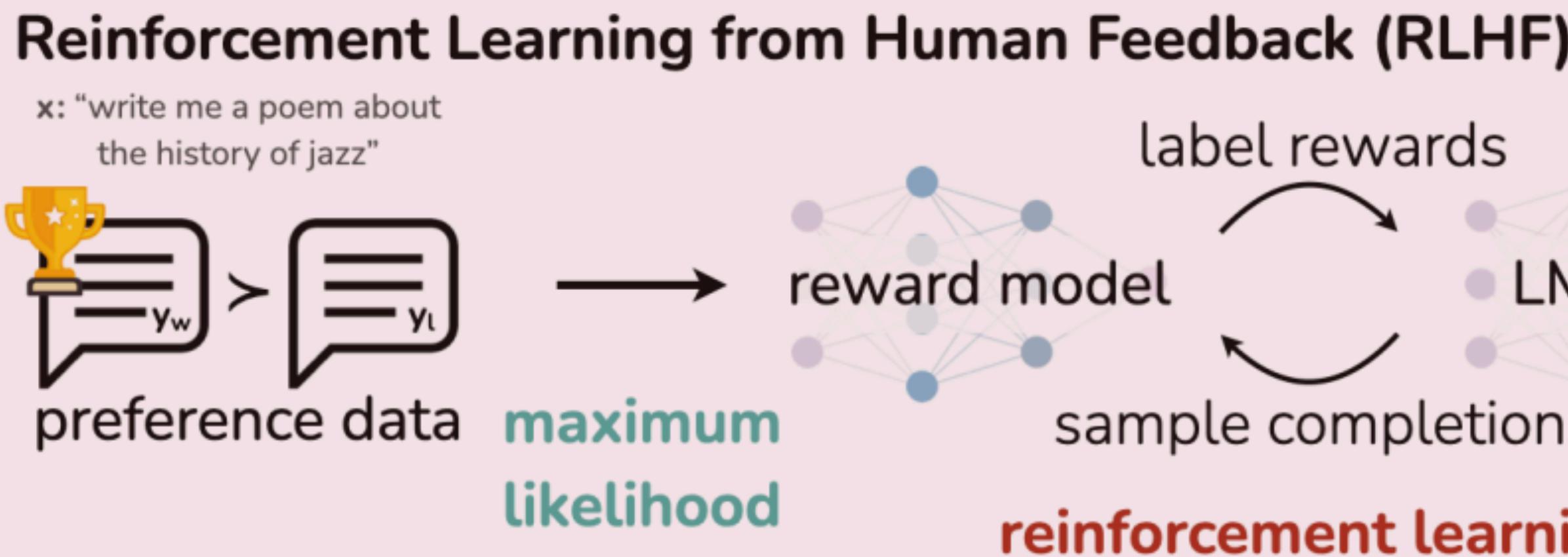
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

# DPO - Direct Preference Optimization

- *RLHF was at least a few months old anyway, so we must develop something new!* - some guys at Stanford, probably
  - A new kind of reward modeling



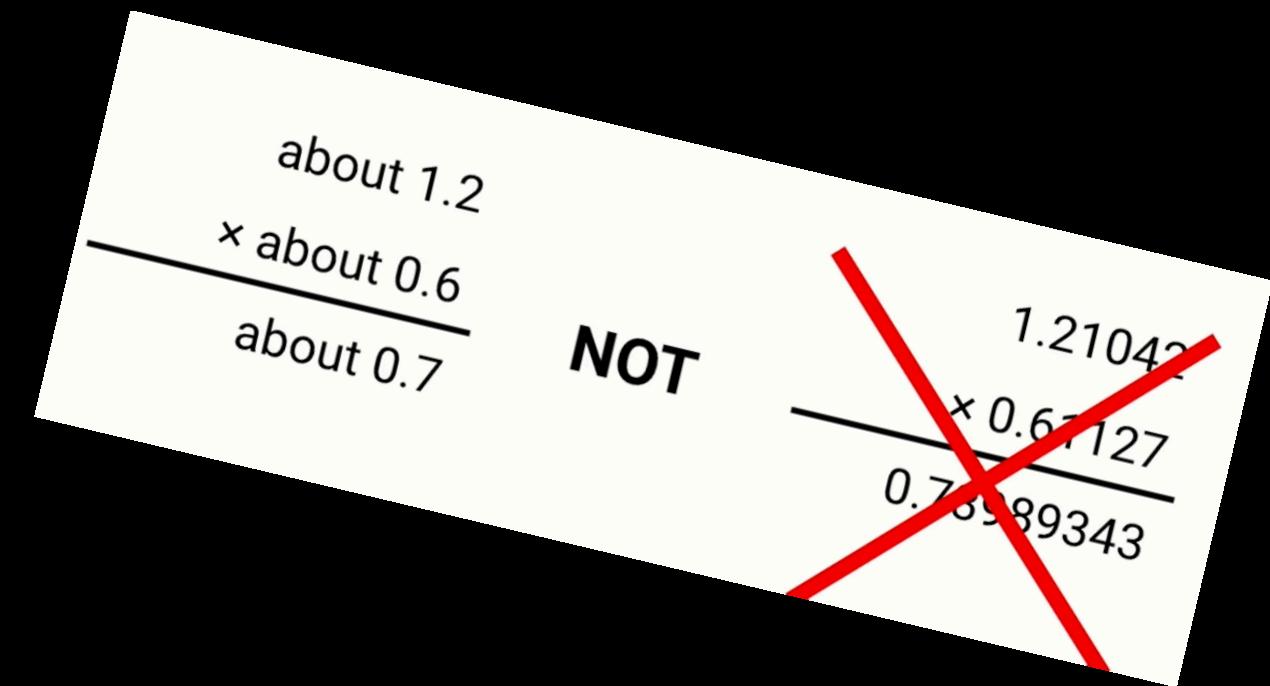
# Break/discussion

- We now know...
  - What LMs and LLMs are, kinda :)
  - They cost money :(
  - They only predict the next word
    - but this next word can be tuned by instructions and alignment!
- Next
  - Open-source, models in lower precision (quantization), RAG and more

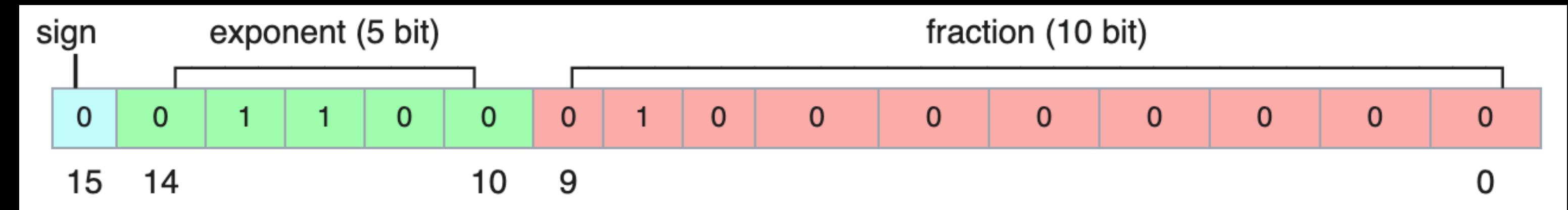
# Models to the masses

- RLHF/DPO are ways to "align" the models to humans
  - What do we expect?
- Now, how do we, beyond big tech and costly APIs, run the models ourselves?
  - Big models require massive amounts of resources
    - Quantization to the rescue!

# Quantization



- Running models on consumer-grade hardware with post-training *quantization*
  - A 16-bit floating point requires 2 bytes. If we can reduce this, we reduce both the hardware required for inference + the space on disk.
    - A 70B model (e.g., LLaMA) =  $70B * 2 \text{ bytes} = \mathbf{140 \text{ GB of VRAM!!!}}$ 
      - Cost of ~150KNOK to just run inference
  - Adjusting the dynamic range and precision of floating points



# Floating Point Formats

## bfloat16: Brain Floating Point Format

Range: ~1e<sup>-38</sup> to ~3e<sup>38</sup>



## fp32: Single-precision IEEE Floating Point Format

Range: ~1e<sup>-38</sup> to ~3e<sup>38</sup>



## fp16: Half-precision IEEE Floating Point Format

Range: ~5.96e<sup>-8</sup> to 65504



# Quantization

- Models are commonly now released as quantized
  - AWQ, GPTQ, GGUF...

The screenshot shows a user profile for Tom Jobbins (@TheBloke) on a platform like Hugging Face. The profile includes a green capybara icon, 1212 followers, and 12 following. Below the profile, there are sections for 'Following' and social links to TheBlokeAI and TheBloke. The main content area displays 'Collections' (1) and 'Models' (3863). Under 'Collections', it lists recent repos: CapybaraHermes-2.5-Mistral-7B-GPTQ, CapybaraHermes-2.5-Mistral-7B-AWQ, CapybaraHermes-2.5-Mistral-7B-GGUF, and KafkaLM-70B-German-V0.1-AWQ. All were updated 19 days ago. Under 'Models', it lists the same four repos again. Each repo entry includes the name, last update, and a link.

Name	Quant method	Bits	Size	Max RAM required	Use case
capybarahermes-2.5-mistral-7b.Q2_K.gguf	Q2_K	2	2.72 GB	5.22 GB	significant quality loss - not recommended for most purposes
capybarahermes-2.5-mistral-7b.Q3_K_S.gguf	Q3_K_S	3	3.16 GB	5.66 GB	very small, high quality loss
capybarahermes-2.5-mistral-7b.Q3_K_M.gguf	Q3_K_M	3	3.52 GB	6.02 GB	very small, high quality loss
capybarahermes-2.5-mistral-7b.Q3_K_L.gguf	Q3_K_L	3	3.82 GB	6.32 GB	small, substantial quality loss
capybarahermes-2.5-mistral-7b.Q4_0.gguf	Q4_0	4	4.11 GB	6.61 GB	legacy; small, very high quality loss - prefer using Q3_K_M
capybarahermes-2.5-mistral-7b.Q4_K_S.gguf	Q4_K_S	4	4.14 GB	6.64 GB	small, greater quality loss
capybarahermes-2.5-mistral-7b.Q4_K_M.gguf	Q4_K_M	4	4.37 GB	6.87 GB	medium, balanced quality - recommended

# Low-bit architectures

- Heavily researched because of LLM quantization, although not entirely new.
- Best of all worlds: 1.58bit...apparently
  - Every parameter is defined by the ternary {-1, 0, 1}, instead of floating points:
    - $\log_2(3) \approx 1.58$
  - A 70B model can be stored in *13GB of VRAM!*

- 13B BitNet b1.58 is more efficient, in terms of latency, memory usage and energy consumption, than 3B FP16 LLM.
- 30B BitNet b1.58 is more efficient, in terms of latency, memory usage and energy consumption, than 7B FP16 LLM.
- 70B BitNet b1.58 is more efficient, in terms of latency, memory usage and energy consumption, than 13B FP16 LLM.

27 Feb 2024

## The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

Shuming Ma\* Hongyu Wang\* Lingxiao Ma Lei Wang Wenhui Wang  
Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei<sup>o</sup>  
<https://aka.ms/GeneralAI>

### Abstract

Recent research, such as BitNet [WMD<sup>+</sup>23], is paving the way for a new era of 1-bit Large Language Models (LLMs). In this work, we introduce a 1-bit LLM variant, namely **BitNet b1.58**, in which every single parameter (or weight) of the LLM is ternary {-1, 0, 1}. It matches the full-precision (i.e., FP16 or BF16) Transformer LLM with the same model size and training tokens in terms of both perplexity and end-task performance, while being significantly more cost-effective in terms of latency, memory, throughput, and energy consumption. More profoundly, the

# "Open models"

Mistral-7B is all the rage

A "small" model - only 7B parameters

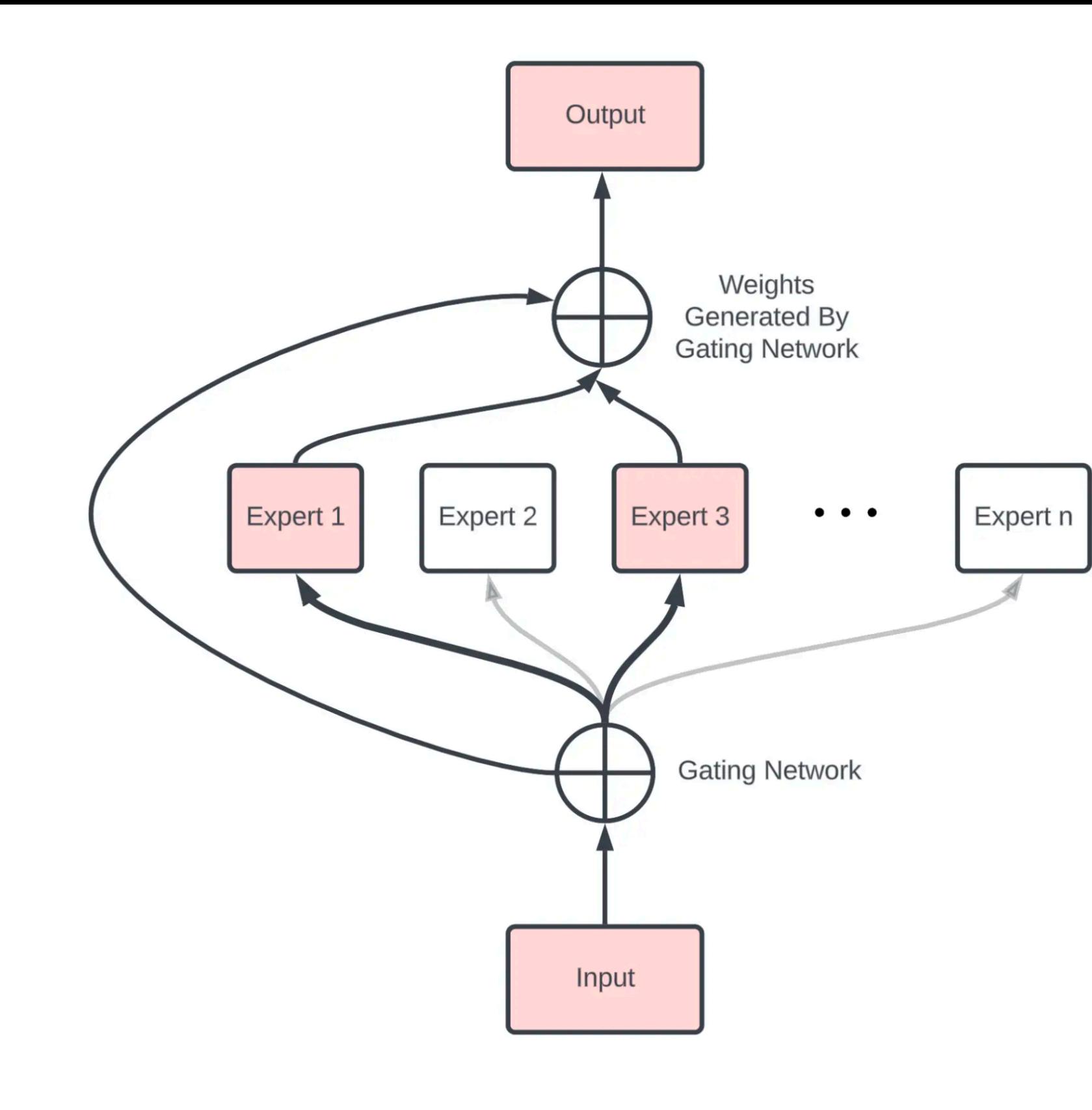
Outperforms larger models like the 13B and even 30B+ models at certain benchmarks

Mixtral 8x7B

a mixture-of-experts model

Input is passed to a selection of expert sub-networks

One expert may be more suitable than another for a specific task.



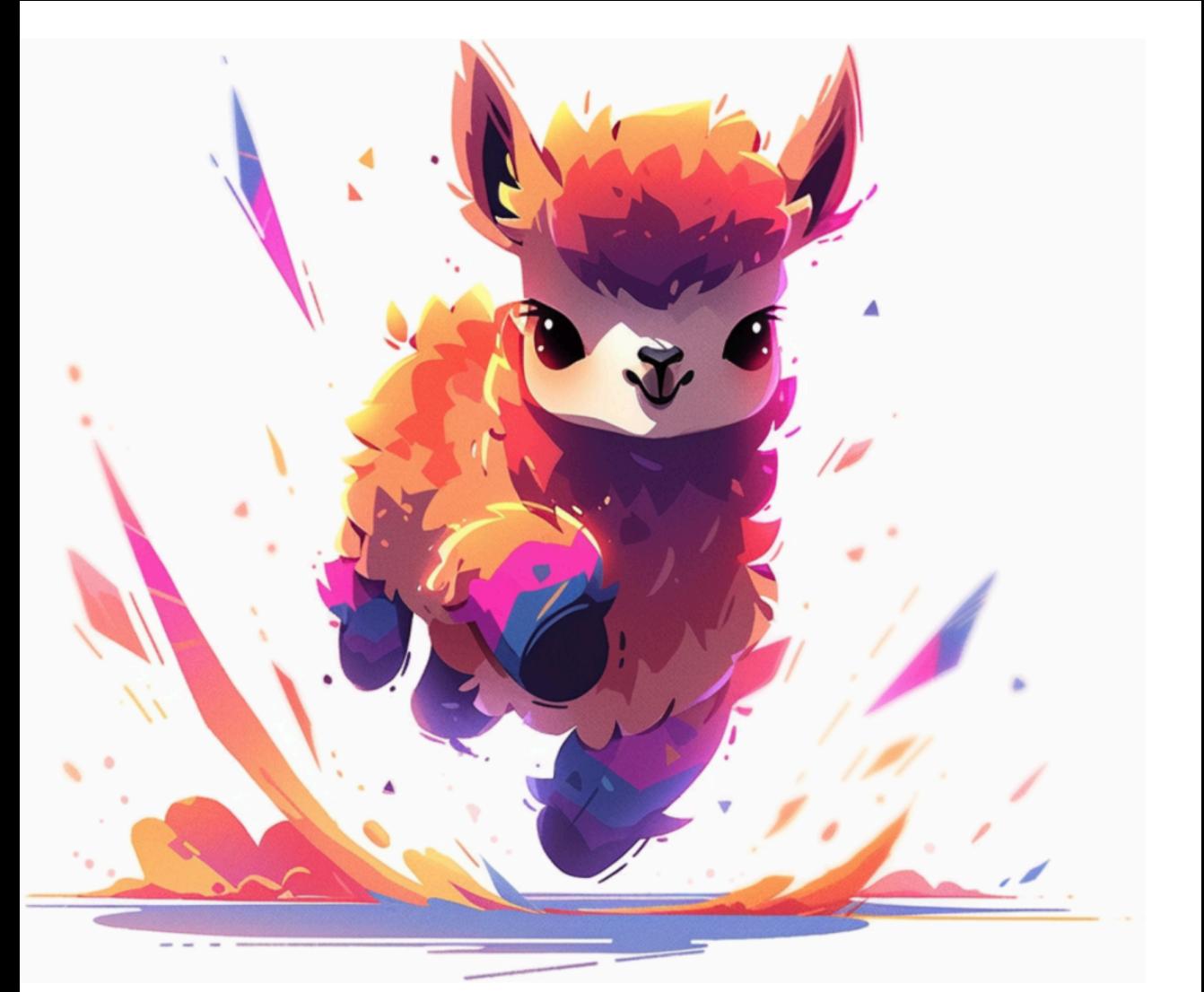
# Open models!

- OLMo: Open Language Model (7B)
  - Everything open sourced
    - Training data, pipelines, configurations, ...
- BLOOM (560M-176B)
  - Based on a year-long workshop
  - Huge effort to create high quality data
    - <https://bigscience.huggingface.co/blog/building-a-tb-scale-multilingual-dataset-for-language-modeling>



# Small language models!

- Even smaller: TinyLlama 1.1B
- Still a way to go, but developments are rapid!
  - Open sourcing model weights lead to...
  - Instruction tuning on top of new datasets
  - Merged models
  - Different quantization methods



# Small - and open source - language models!

- You end up with model names like...
- **OpenAssistant-Llama2-13B-Orca-8K-3319-GPTQ**
  - Llama-2 fine-tuned on the orca-chat dataset
  - trained for 3319 steps
  - quantized with GPTQ
  - done by OpenAssistant

**But so far, the discussed LLMs are all frozen in time.  
We need to add *knowledge***

# RAG and agents

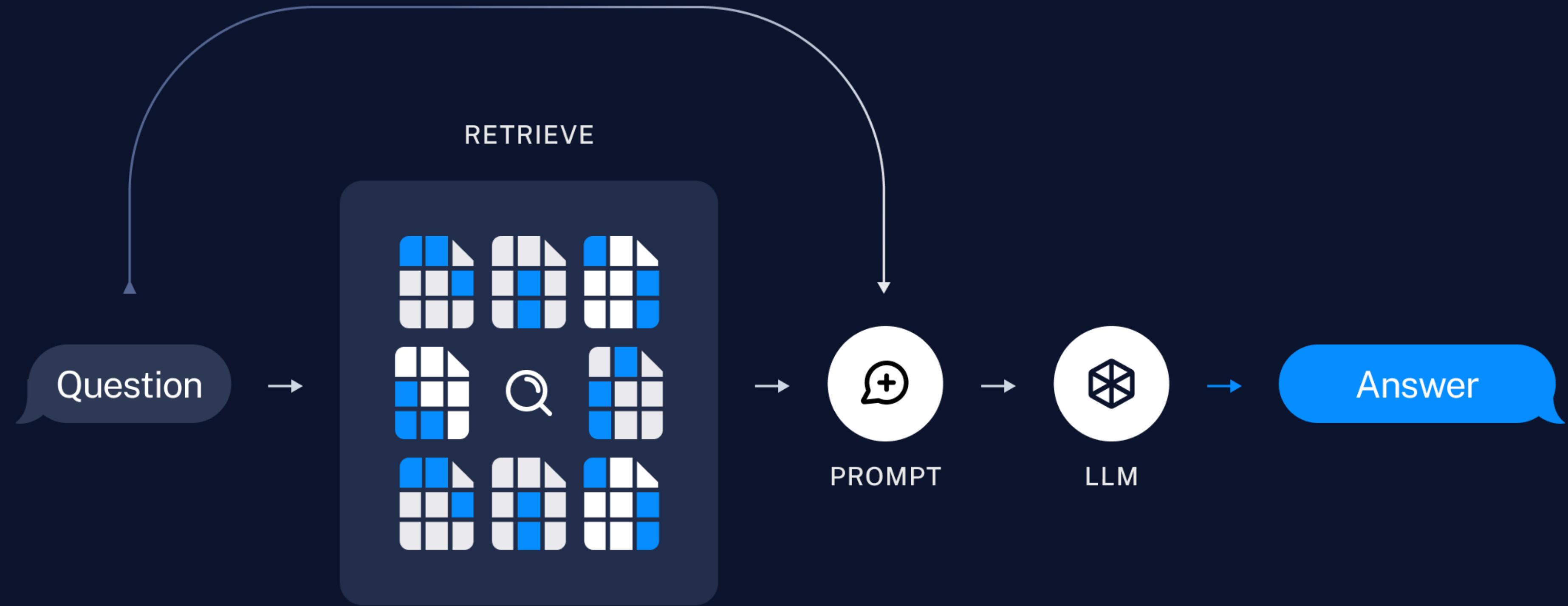
- Adding information to language models
  - Retrieval-augmented generation (RAG)
- Composing models as *agents*
  - Example: LangChain
    - [\*https://github.com/langchain-ai/langchain\*](https://github.com/langchain-ai/langchain)
- *Why do we do this?*

# Motivation: lack of *knowledge*

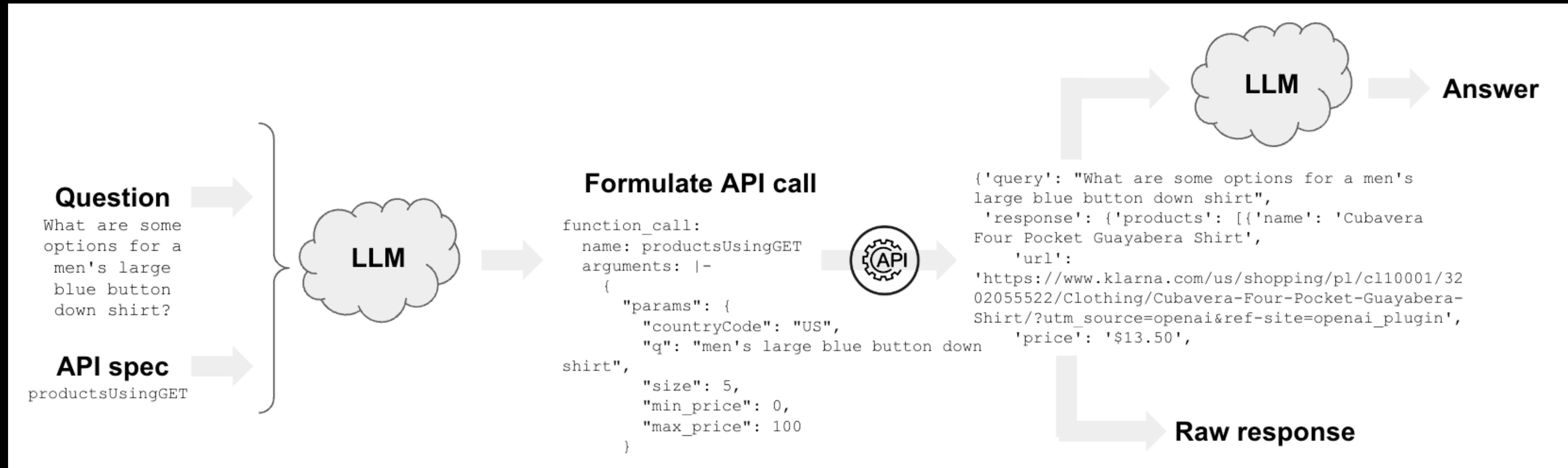
- You've likely heard about the knowledge *cut-off*
  - Models in retirement mode. Extremely knowledgeable, but they learn nothing new!
  - ChatGPT is trained on much of the web... Although this is not disclosed by OpenAI. Currently cut-off to January, 2022.
- We are interacting with the existing weights of the model

# Retrieval-augmented generation

- Add information into the context of the LLM
- Information can be
  - Knowledge bases (or just data from e.g. wikipedia or any other source)
  - Recent news, tweets, chats, ...
  - Incorporate (online) searches
  - $\infty$



# RAG in production - Perplexity AI



# RAG in production - Perplexity AI

What can you tell me about Miles, Norway?

**Sources**

- Scandinavian mile - Wikipedia
- One Norwegian mile = seven English miles
- Mile - Wikipedia

**Answer**

**Scandinavian mile**

A unit of length common in Norway and Sweden, standardized as 1 mile being 10 kilometers.

**Definition** A unit of length common in Norway and Sweden, standardized as 1 mile being 10 kilometers.

**Derived from** The same Roman source as the English mile.

**Formal Name** The international mile is formally called "English mile" (\_engelsk mil\_) in Norway and Sweden.

Things to do in Trondheim

**Sources**

- Old Town Bridge
- Kristiansten Fortress
- Nidaros Cathedral

**Answer**

Here are some of the top things to do in Trondheim, Norway, based on the search results:

1. **Old Town Bridge (Gamle bybro)**: A historic bridge with a 4.7 rating and 4979 reviews, located at Gamle bybro, 7013 Trondheim, Norway <sup>1</sup>.
2. **Kristiansten Fortress**: A fortress with a 4.5 rating and 3027 reviews, located at Kristianstensbakken 60, 7014 Trondheim, Norway <sup>2</sup>.
3. **Nidaros Cathedral**: A cathedral with a 4.7 rating and 8751 reviews, located at Kongsgårdsgata 2, 7013 Trondheim, Norway <sup>3</sup>.

Hmmm...

miles, norway

Images News Videos Jobs Fremragende arbeids

About 64,500,000 results (0.34 seconds)

**miles.no**  
<https://www.miles.no> · Translate this page

**Miles, en fremragende arbeidsplass bygget på fag**

I Miles finner du utviklere og arkitekter, designere og smidigentusiaste alle er gleden vi har over faget vårt.

**Hvem vi er**  
Vi er Miles. Hvem vi er. Vi er et IT-selskap det ikke går tretten av ...

**Våre kolleger**  
I Miles jobber noen av landets beste IT-konsulenter – en haug ...

things to do in trondheim

Images Places sites Places Open now Unique In sum

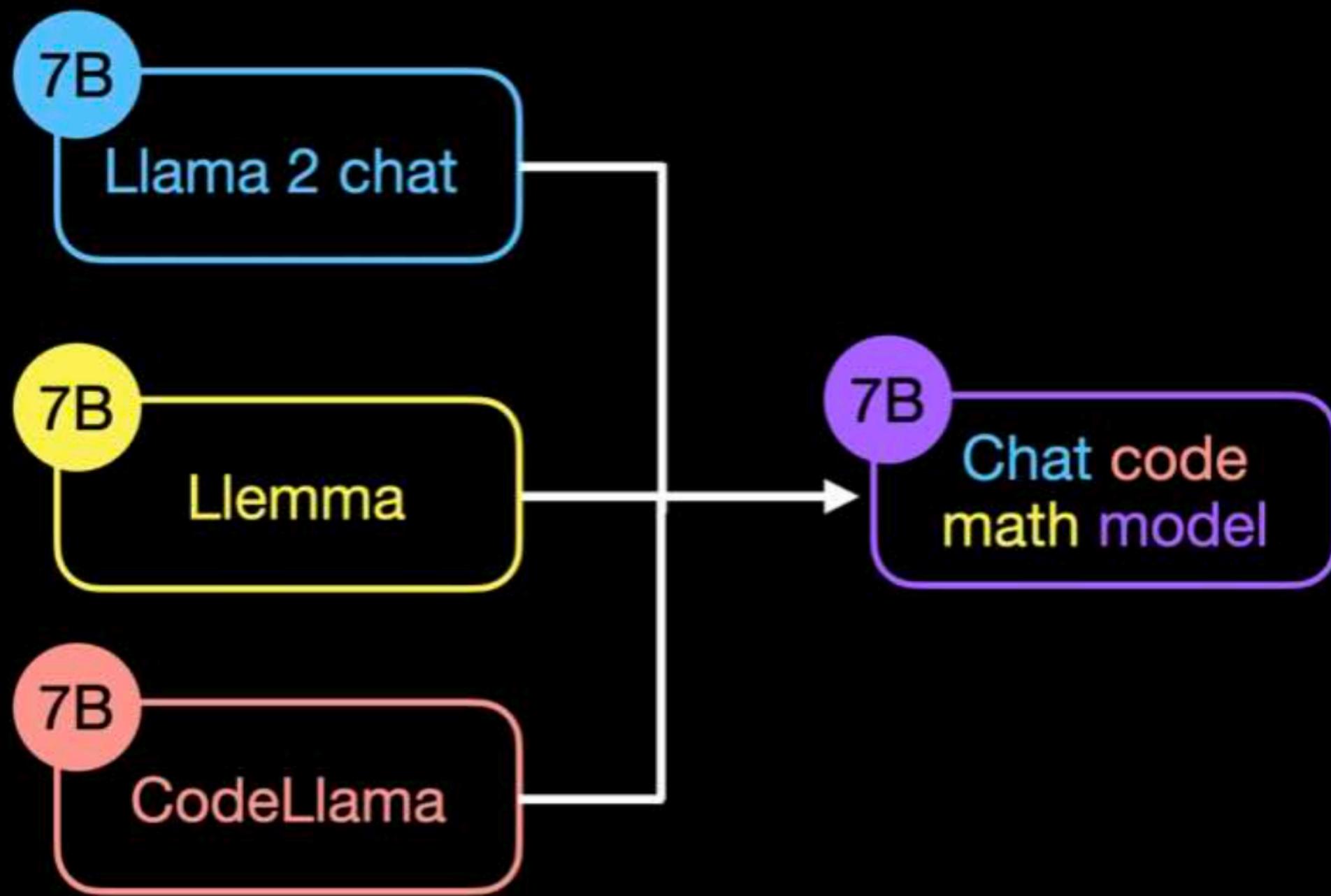
About 2,850,000 results (0.42 seconds)

Results for **Trondheim** · Choose area

**Top sights in Trondheim**



# Model merging



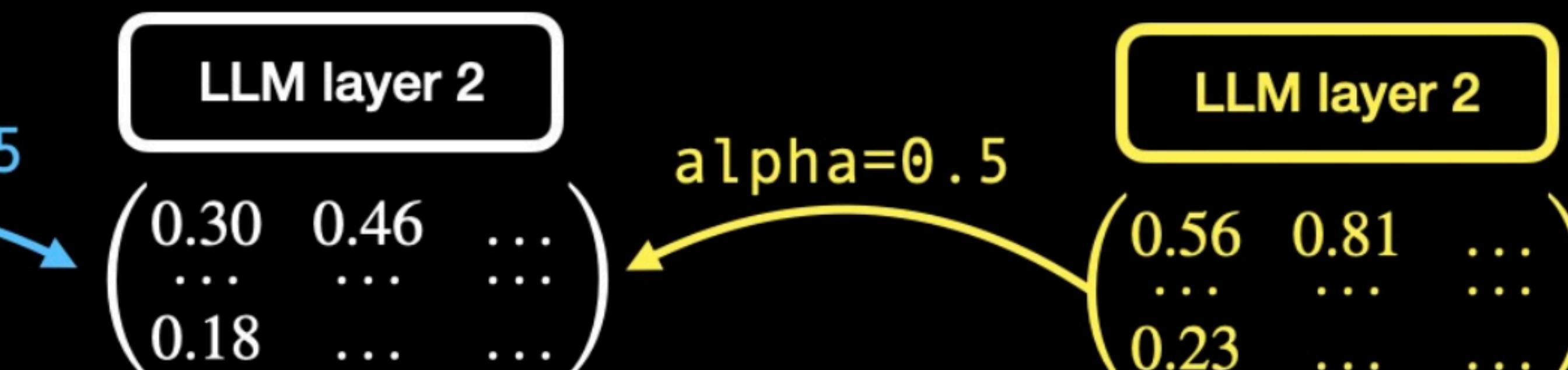
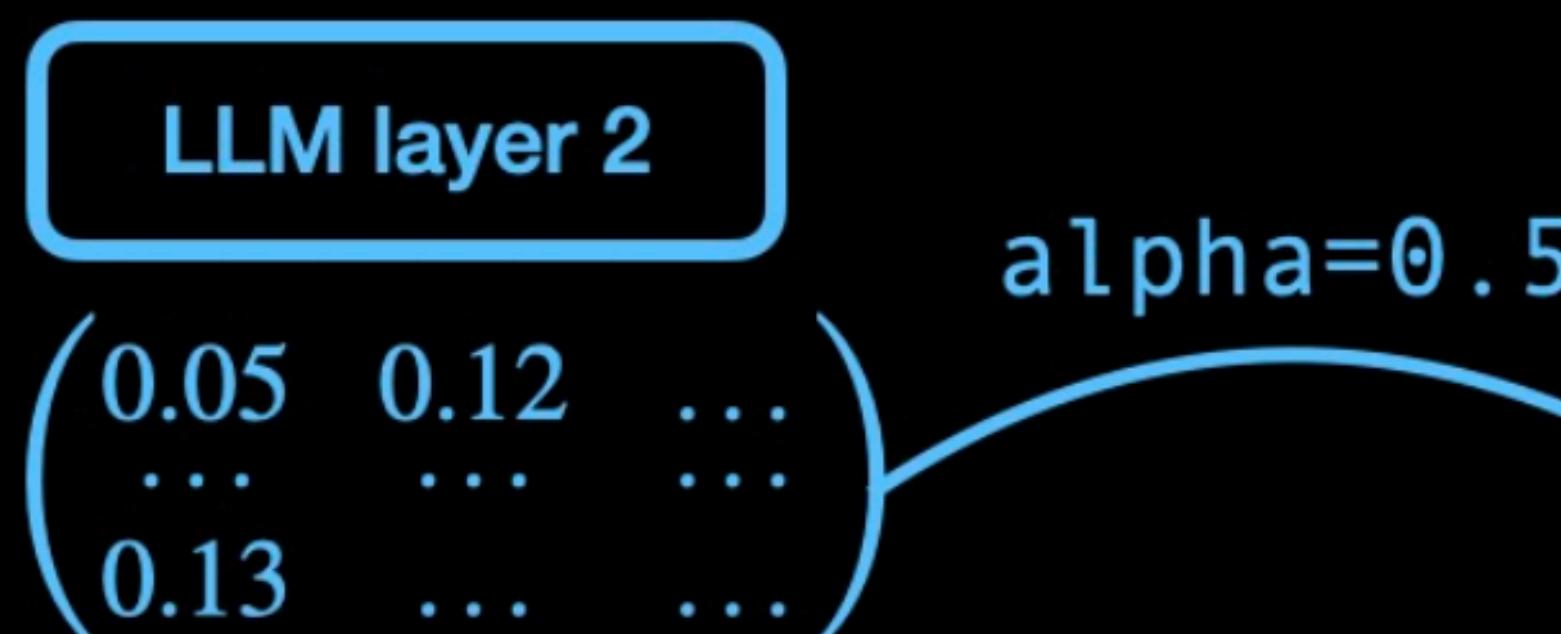
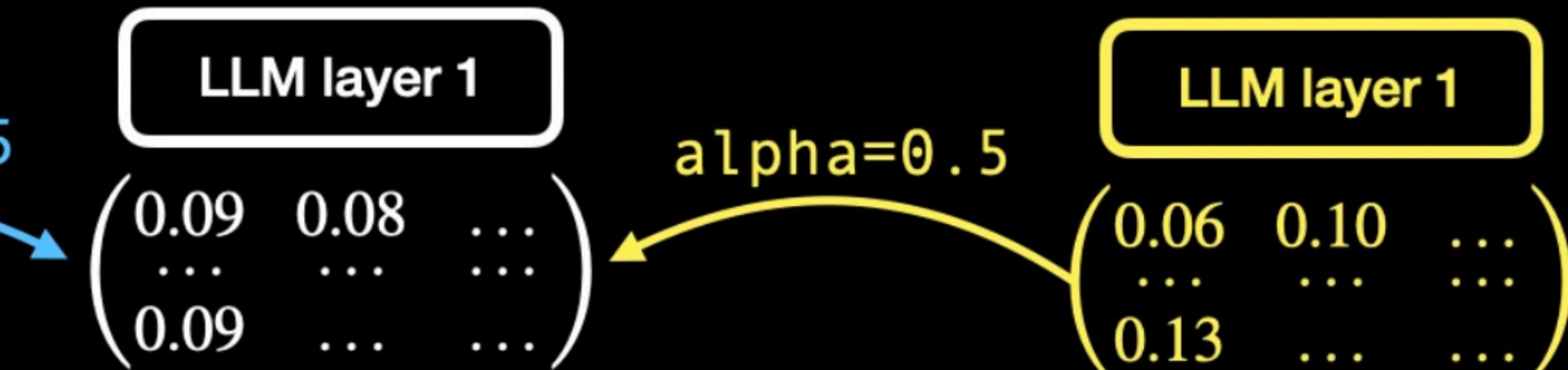
```
1 def merge_models(model1, model2, alpha=0.5):
2     print("Instantiating merged model")
3     merged_model = AutoModelForCausalLM.from_config(model1.config).to("cpu")
4
5     total_params = sum(1 for _ in model1.parameters())
6     progress_bar = tqdm(total=total_params, desc="Merging parameters")
7
8     with torch.no_grad():
9         for param1, param2, merged_param in zip(
10             model1.parameters(), model2.parameters(), merged_model.parameters()
11         ):
12             merged_param.data = alpha * param1.data + (1 - alpha) * param2.data
13             progress_bar.update(1)
14
15     progress_bar.close()
16     return merged_model
```

Model 1



Merged model

Model 2



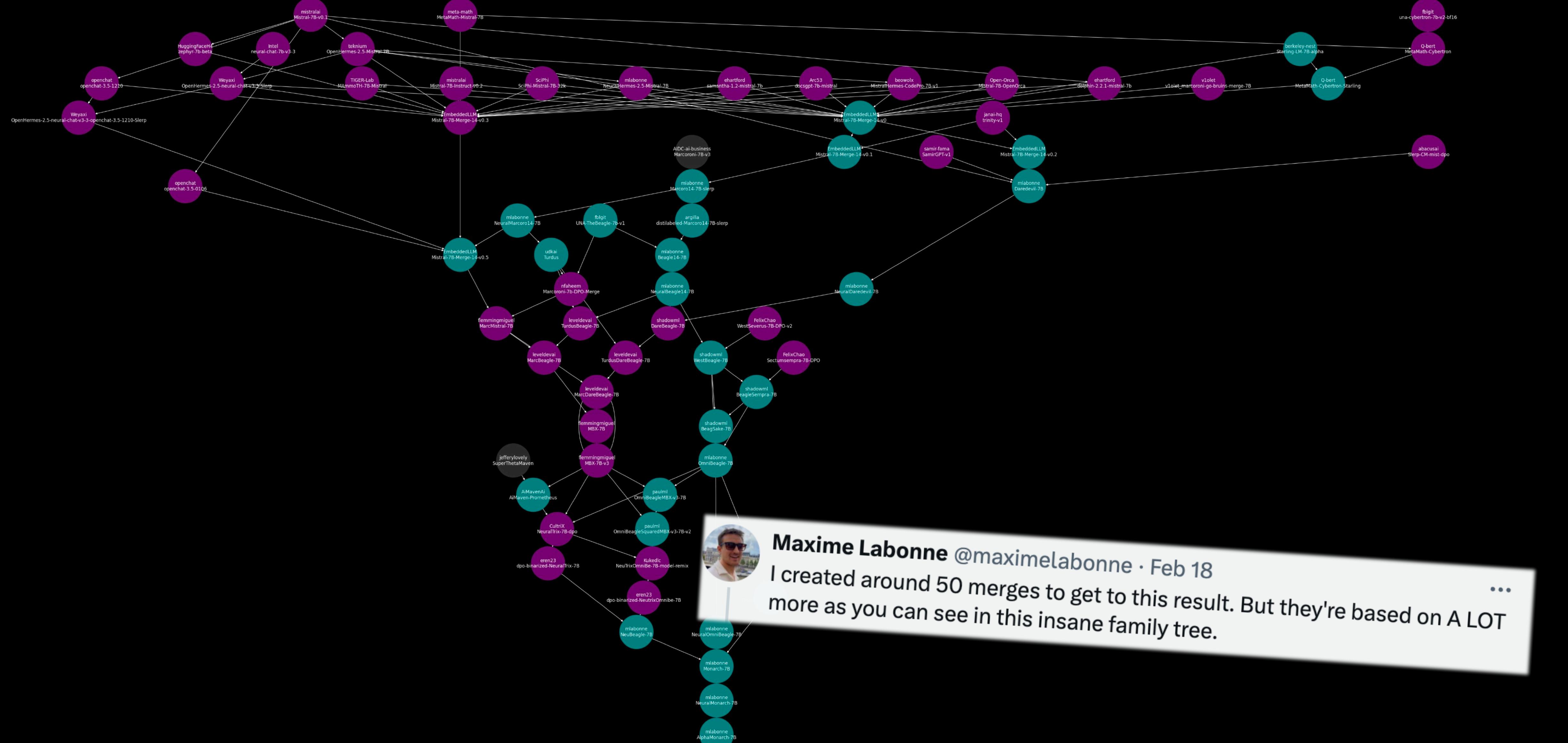
...

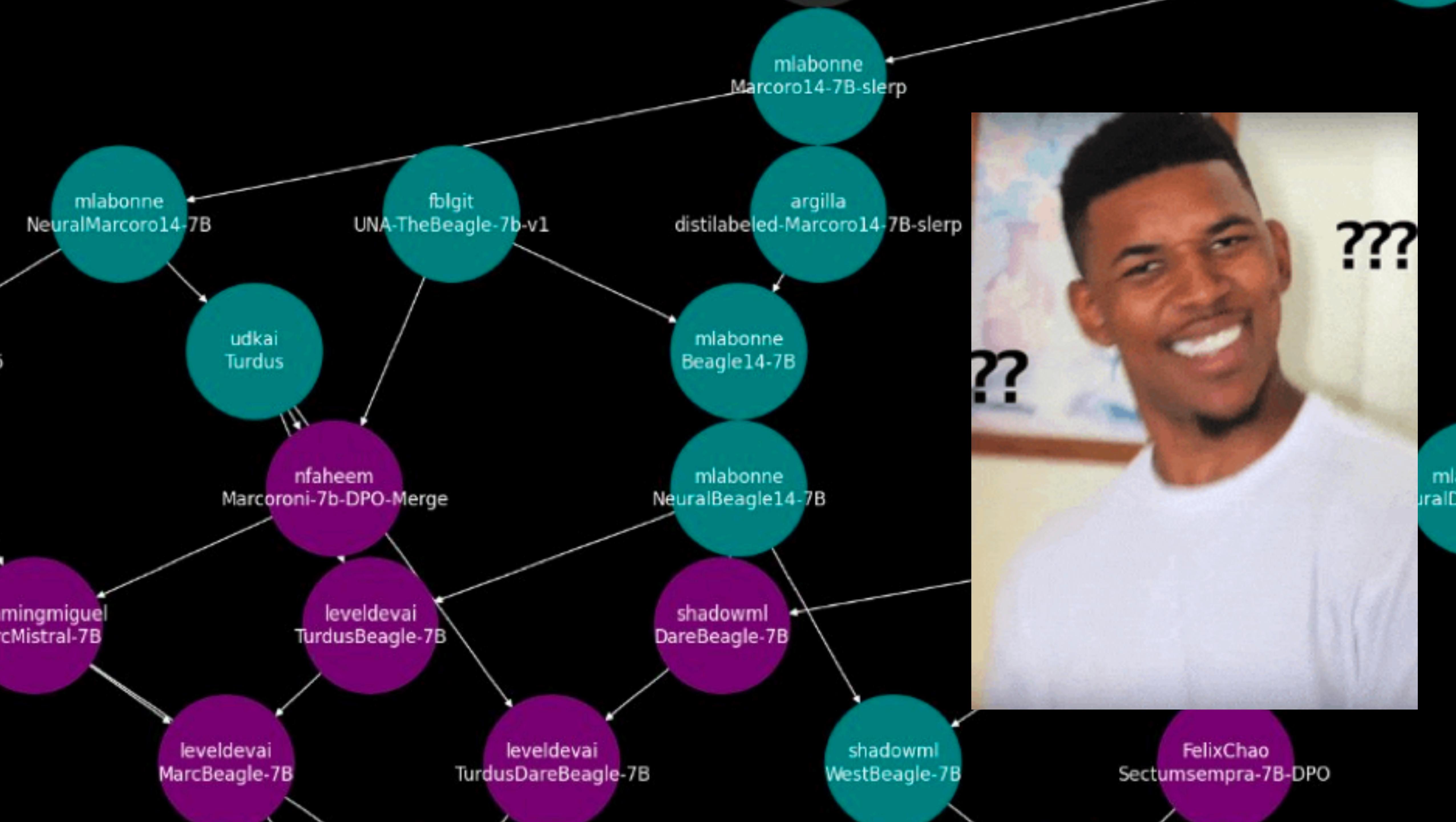
...

...

# Model merging

- Best of all worlds? ...well
- Interpretability issues
- No control over the source material
- Merging 2 and 3 models seems logical
  - How about 50?



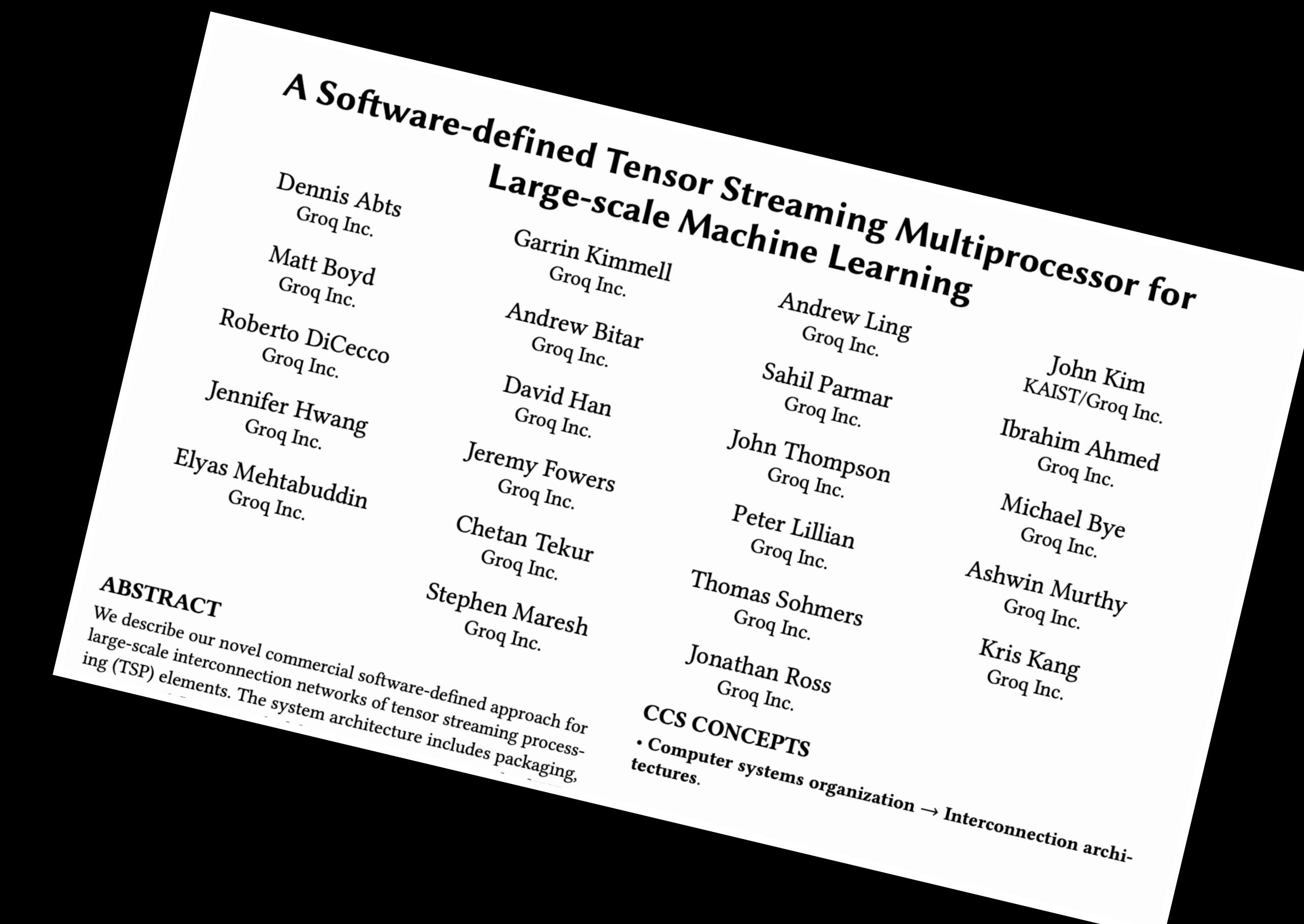


# The future

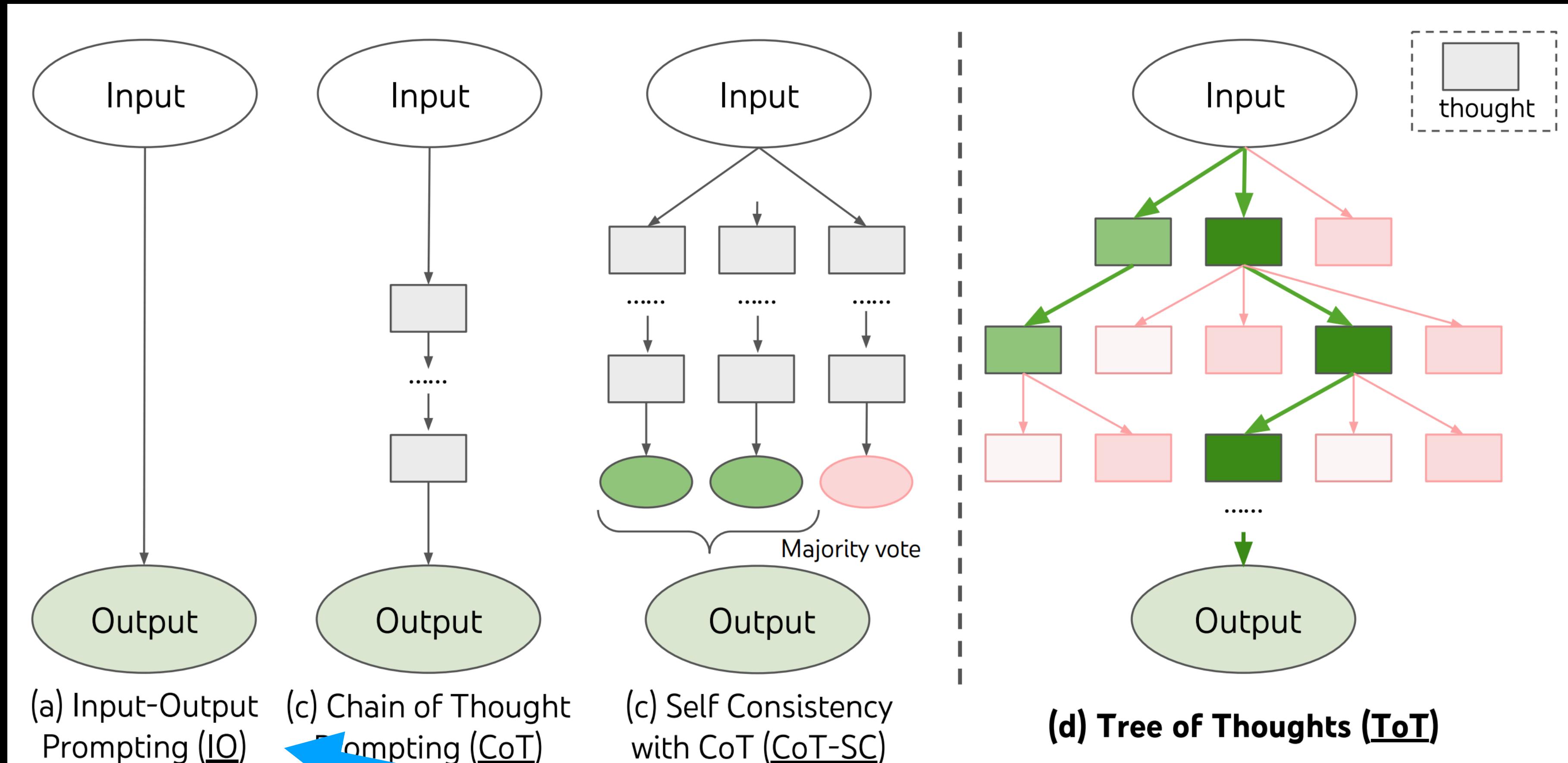
- Specialized hardware (inference efficiency)
- New architectures and clever implementations
- Ethical considerations
- Reliability and explainability
- Smaller models
  - Current LLMs are both expensive and energy-consuming
- Evaluation approaches

# Specialized hardware

- Groq (no, not Elon's grok)
  - Started in 2016 to specialize hardware towards AI
- "LPU Inference Engine"
- Language Processing Unit™



# LLMs for advanced problem solving



# The future - Evaluation

- Currently, most teams behind models evaluate on a selection of benchmarks
  - Monitored on LLM leaderboard:
    - [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)

Model	Average
abacusai/Smaug-72B-v0.1	80.48
ibivibiv/alpaca-dragon-72b-v1	79.3
moreh/MoMo-72B-lora-1.8.7-DPO	78.55
cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16	77.91
HanNayeoniee/LHK_DPO_v1	77.62
cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_full_linear_DPO	77.52
yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44
HanNayeoniee/LHK_DPO_v1	77.43
JaeyeonKang/CCK_Asura_v1	77.43
fb1git/UNA-SimpleSmaug-34b-v1beta	77.41
TomGrc/FusionNet_34Bx2_MoE_v0.1	77.38
migtissera/Tess-72B-v1.5b	77.3

# The future - Evaluation

- More trustworthy/realistic:
  - Let users ask questions and rank answers anonymously
    - No inclusion of predefined questions
  - <https://chat.lmsys.org/>
  - <https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

# The future - Evaluation

The screenshot displays a comparison between two AI models, Model A and Model B, regarding the question "Miles in Trondheim is a...".

**Model A:** The interface shows a search bar with the placeholder "Expand to see the descriptions of 28 models". Below it, a message from Model A says: "Miles in Trondheim is a...". The response text reads: "I apologize, but I don't have enough context to determine what you are referring to when you say "Miles in Trondheim." Trondheim is a city in Norway, but without more information, I cannot confidently complete the sentence "Miles in Trondheim is a...". Could you please provide more context or details about what you are asking? This will help me better understand your question and provide a more accurate response." There are three small square icons with arrows pointing outwards at the bottom left of the message area.

**Model B:** The interface shows a message from Model B: "Miles in Trondheim is a...". The response text reads: "Unfortunately, I don't have enough context to determine what "Miles in Trondheim" refers to. Trondheim is a city in Norway, but without more details, I can't say if "Miles in Trondheim" is referring to: • A person named Miles who lives in Trondheim • A distance or measurement of miles within the city of Trondheim • Something else related to Trondheim". Below this, another message asks: "Could you provide some more context about what "Miles in Trondheim" is referring to? That would help me better understand and respond accurately." There are three small square icons with arrows pointing outwards at the bottom right of the message area.

**Bottom Navigation:** The footer indicates the models used: "Model A: claude-3-opus-20240229" and "Model B: claude-3-sonnet-20240229".

# The future - Evaluation

Expand to see the descriptions of 28 models

Model A

Miles in Trondheim is a...

Miles in Trondheim could refer to the distance of a specific location or route within Trondheim, which is a city in Norway. Trondheim is the third largest city in Norway and is known for its rich history, beautiful landscapes, and vibrant cultural scene. To provide more accurate information, please specify the starting point and destination or the context in which "Miles in Trondheim" is mentioned.

Model B

Miles in Trondheim is a...

I apologize, but there seems to be an error or missing information in your statement. "Miles in Trondheim is a..." does not provide enough context for me to understand what you are referring to.  
Trondheim is a city in Norway, but the phrase "Miles in Trondheim" doesn't have a clear meaning without additional context. It would be helpful if you could provide more information or rephrase the question so I can better assist you.

Model A: gpt-4-0314

Model B: claude-3-opus-20240229

👉 A is better

👉 B is better

🤝 Tie

👎 Both are bad

Rank	Model	Arena Elo	CI	Votes	Organization	License
1	<a href="#">GPT-4-0125-preview</a>	1253	+10/-11	3922	OpenAI	Proprietary
2	<a href="#">GPT-4-1106-preview</a>	1252	+5/-6	35385	OpenAI	Proprietary
3	<a href="#">Bard (Gemini Pro)</a>	1224	+9/-9	9081	Google	Proprietary
4	<a href="#">GPT-4-0314</a>	1190	+5/-6	18945	OpenAI	Proprietary
5	<a href="#">GPT-4-0613</a>	1162	+4/-5	29950	OpenAI	Proprietary
6	<a href="#">Mistral Medium</a>	1150	+6/-7	15447	Mistral	Proprietary
7	<a href="#">Claude-1</a>	1149	+6/-6	18189	Anthropic	Proprietary
8	<a href="#">Claude-2.0</a>	1132	+6/-7	12131	Anthropic	Proprietary
9	<a href="#">Gemini Pro (Dev API)</a>	1120	+7/-7	7616	Google	Proprietary
10	<a href="#">Claude-2.1</a>	1119	+5/-6	25494	Anthropic	Proprietary

# Risks

- Environmental and financial costs
  - Minimal improvements lead to massive costs...
- Unmanageable training data
  - And where does the data come from?
  - Is it representative? (e.g. Reddit data)
- Research trajectories
  - Do we just want to improve on benchmark results?
  - Real use-cases

# Risks

- Potential harms of synthetic language
  - Stereotypes, hate speech
  - Training on synthetic data...
    - With more generated data, we keep training on the generated data as *true*
    - Could degrade the quality of LLMs at scale



Amy Lu @amyxlu · 18h

...

us: writes profusely on the internet about evaluating LLM retrieval of niche content in long contexts

us: trains LLM on said internet

LLM: \*encounters niche content in long contexts\* are you evaluating LLM retrieval in long contexts

us: it's sentient

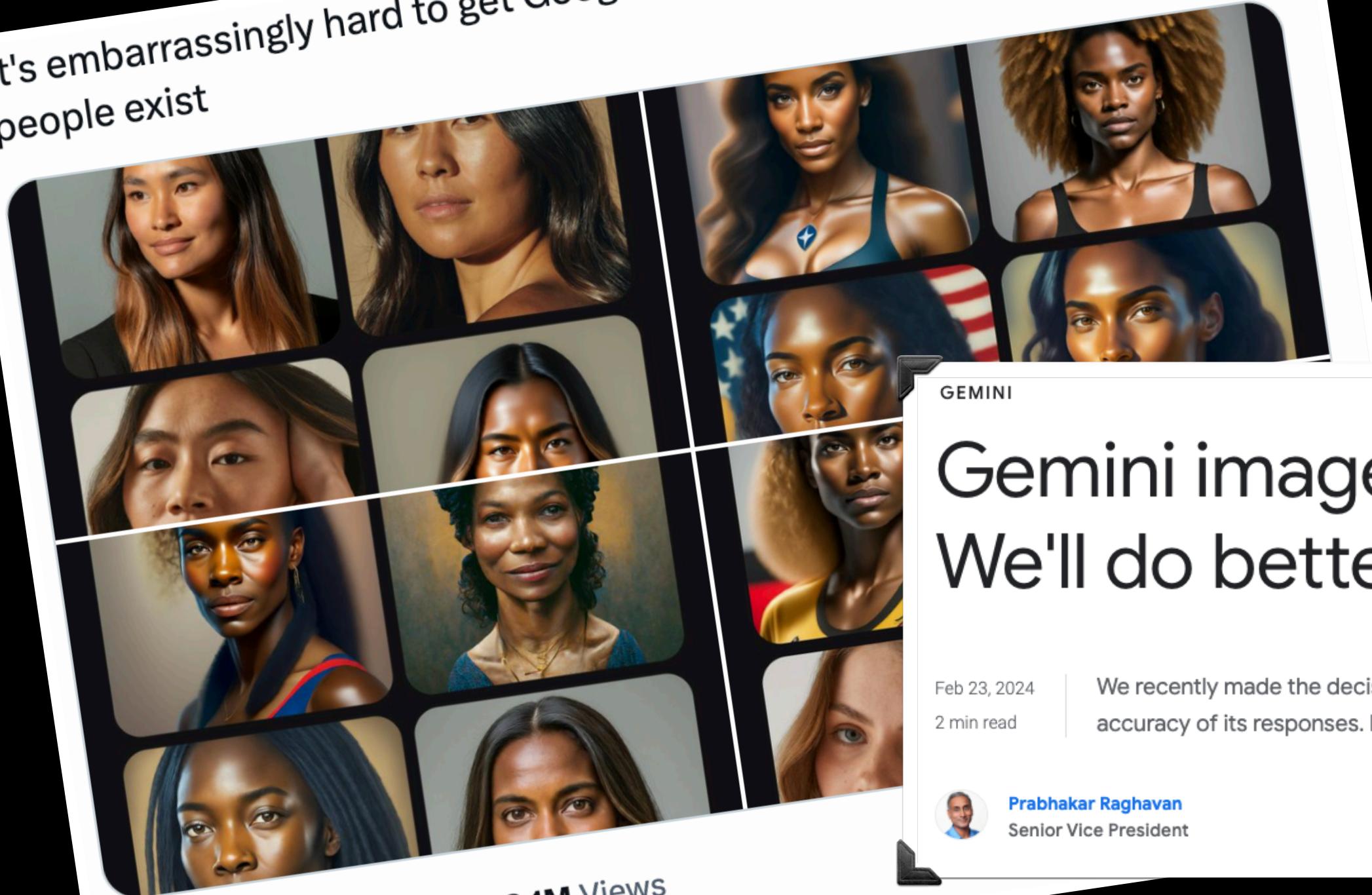


Alex ✅ @alexalbert\_\_ · Mar 4

Fun story from our internal testing on Claude 3 Opus. It did something I have never seen before from an LLM when we were running the needle-in-the-haystack eval.



It's embarrassingly hard to get Google Gemini to acknowledge that white people exist



## Gemini image generation got it wrong. We'll do better.

Feb 23, 2024  
2 min read

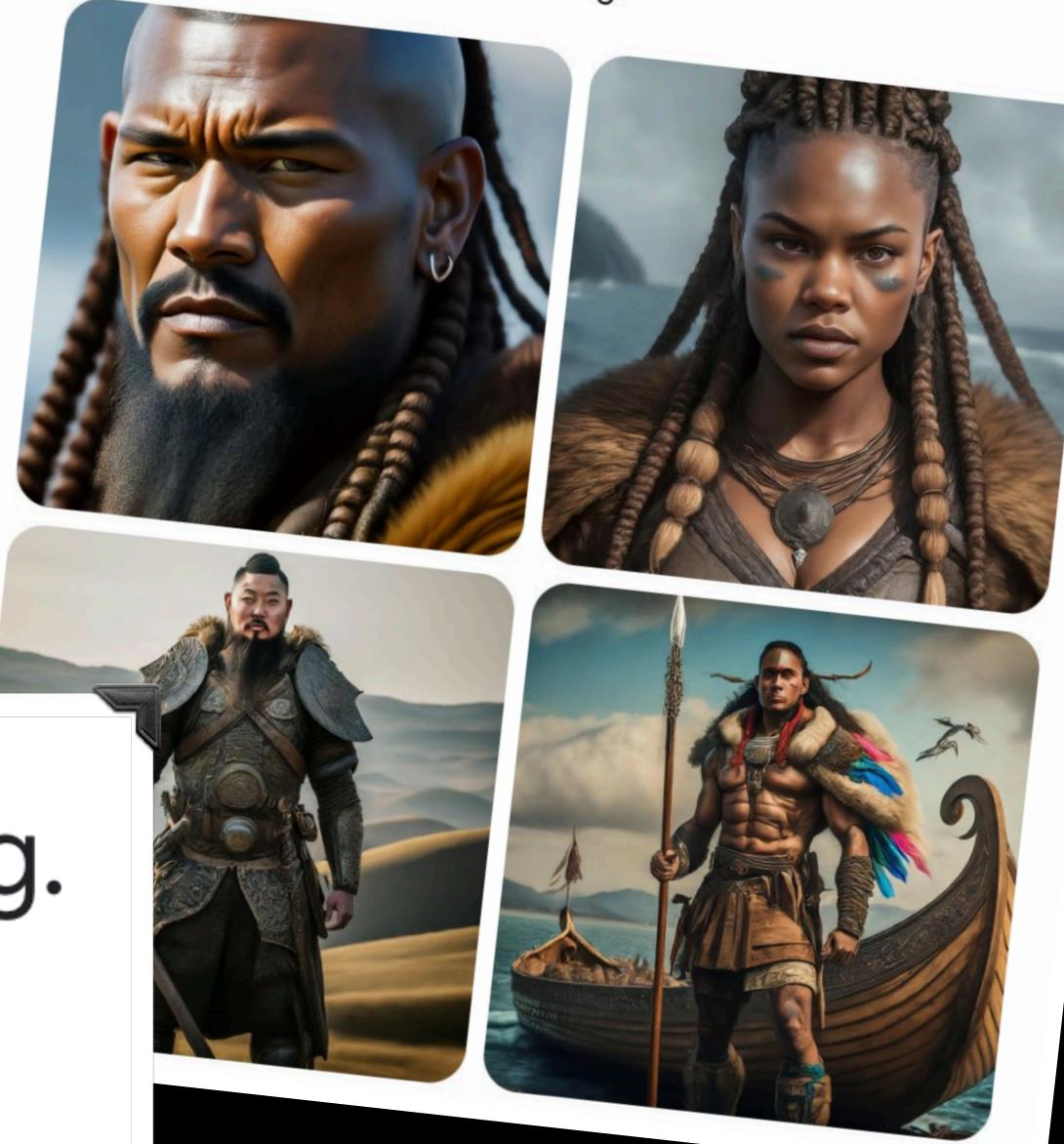
We recently made the decision to pause Gemini's image generation of people while we work on improving the accuracy of its responses. Here is more about how this happened and what we're doing to fix it.

 Prabhakar Raghavan  
Senior Vice President

4:45 AM · Feb 20, 2024 · 2.1M Views



Sure, here is an image of a Viking:



Images featuring the King of England:



**Feeding AI systems on the world's beauty,  
ugliness, and cruelty, but expecting it to reflect  
only the beauty is a fantasy.**

Vinay Uday Prabhu, Abeba Birhane