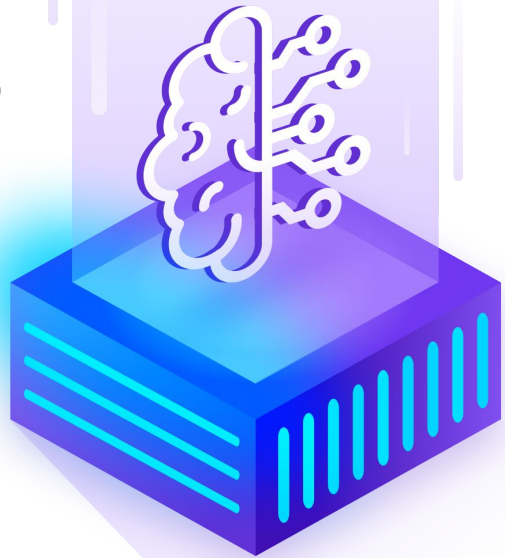
Multimodal LLMs

Scaling Performance



Neha Sharma
Product Manager @Ori

Multimodality is fun and useful

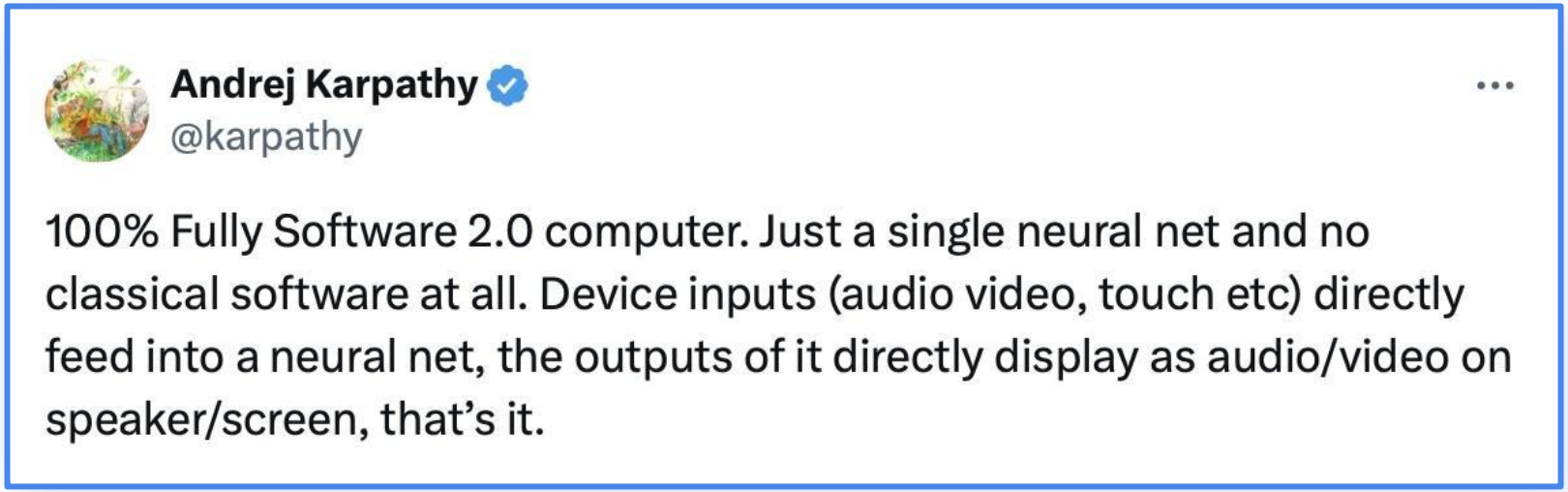
I saw this bird today on my walk. Tell me more about it.

The bird in your photo appears to be a European Robin (Erithacus rubecula). Here are some interesting facts about the European Robin:

1. **Appearance**: The European Robin is easily recognizable by its orange-red breast and face, which contrast with its grey-brown upper parts and white belly. Both male and female robins look very similar.

2. **Habitat**: Robins are common across Europe and can be found in a variety of habitats including gardens, woodlands, and parks. They are also present in parts of western Asia and North

Multimodal LLMs are going to be transformational



Are diffusion models same as multimodal LLMs?

Transformer models	Diffusion models
Autoregressive Decoders	Denoising Autoencoders
Predict related data. E.g. next token, nearby pixel	Predict unrelated data. E.g. far off pixels, or randomised noise in nearby pixels
More granular, slower (token /s)	Less granular, faster (pixels /s)
GPT, Original Dall-E	Dall-E 3

Are diffusion models same as multimodal LLMs?

Left Brain VS Right Brain

Logical

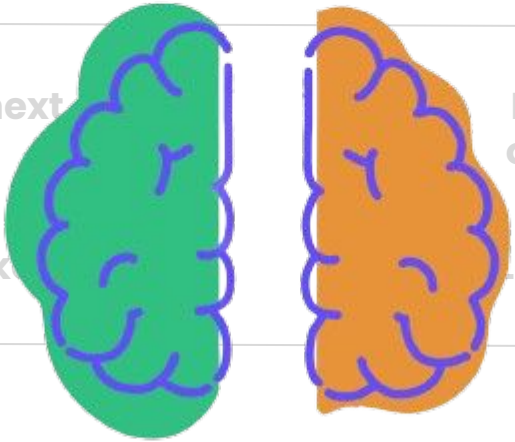
Analytical

Linear

Verbal

Factual

Sequential



Creative

Intuitive

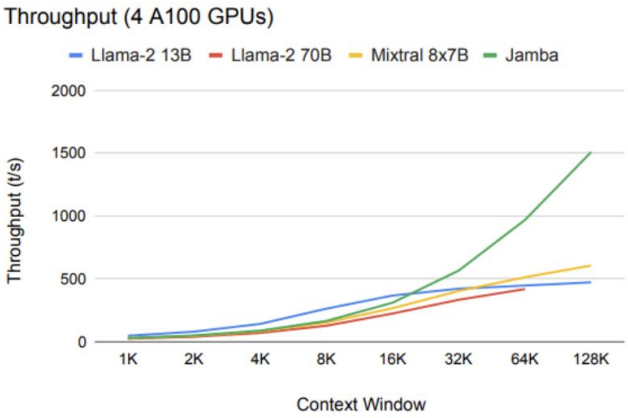
Artistic

Non-verbal

Emotional

Imaginative

There are also non-transformer LLMs



JAMBA = MAMBA + Transformers + MoE

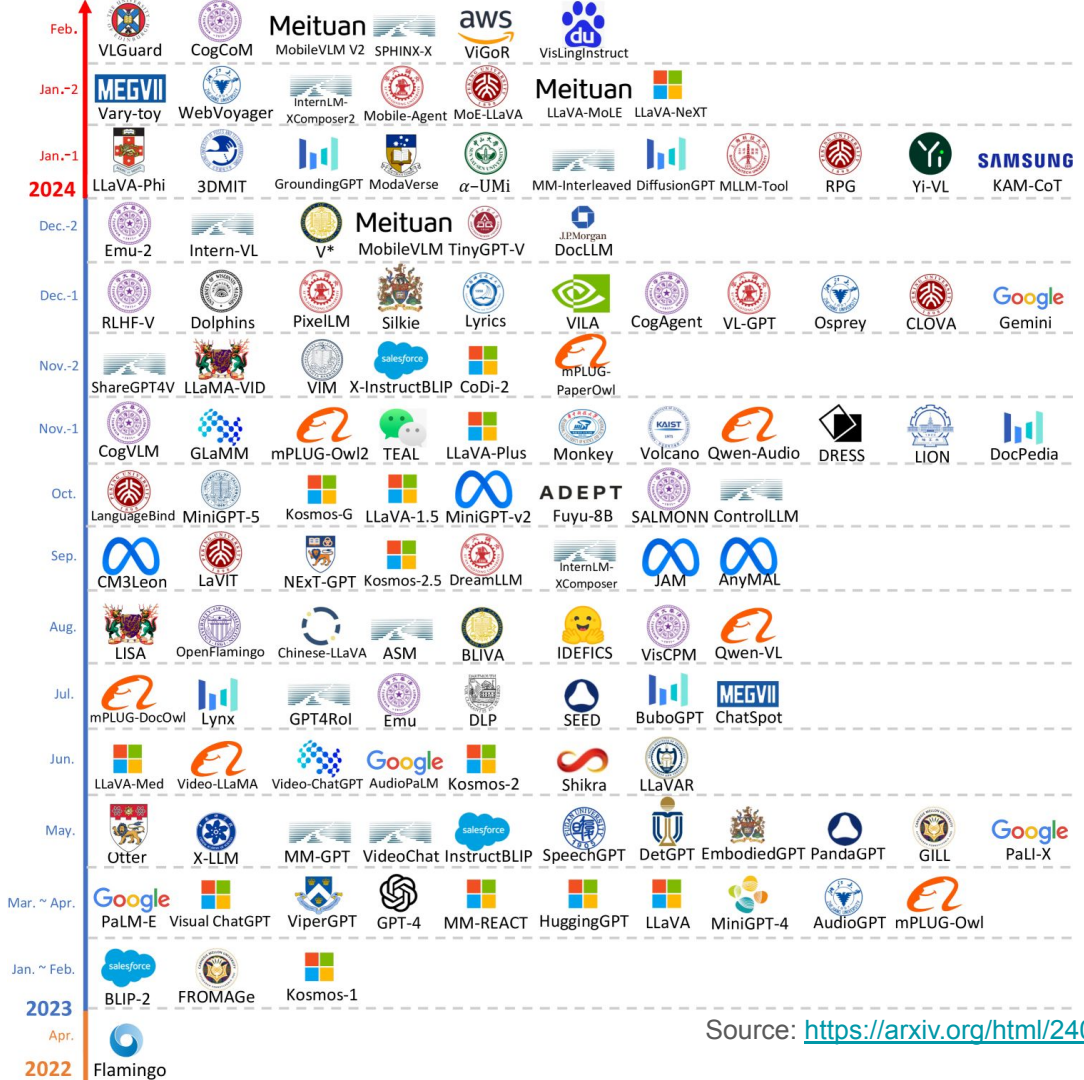
RNNs: hidden state, training cannot be parallelised

Transformers: attention mechanism is quite good but slows down with context window size, since memory scales quadratically with sequence length.

MAMBA: structured state space sequence (S4) model with hidden state like RNN, but can be trained like CNN so training is efficient

Vision State Space Models (SSMs) : Vision Mamba vs VMamba vs S4ND

The rise of multimodal LLMs



Source: https://arxiv.org/html/2401.13601v4

How do multimodal LLMs work?

1

We need a **vision encoder**

2

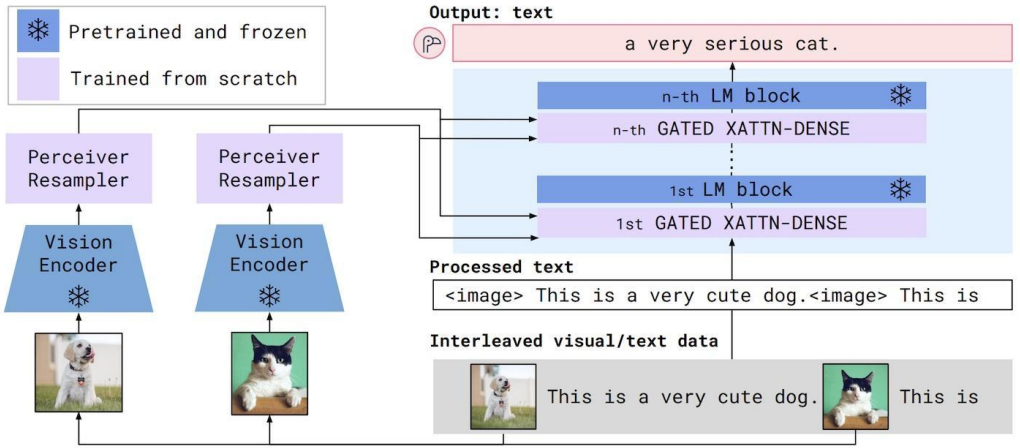
Add some layers to a pre-trained LLM model

3

Do some joint training with images and text pairs

E.g. **CLIP:** Contrastive Language-Image Pre-training

For each image-text pair, the image and text embeddings are close (*cosine similarity*) to each other

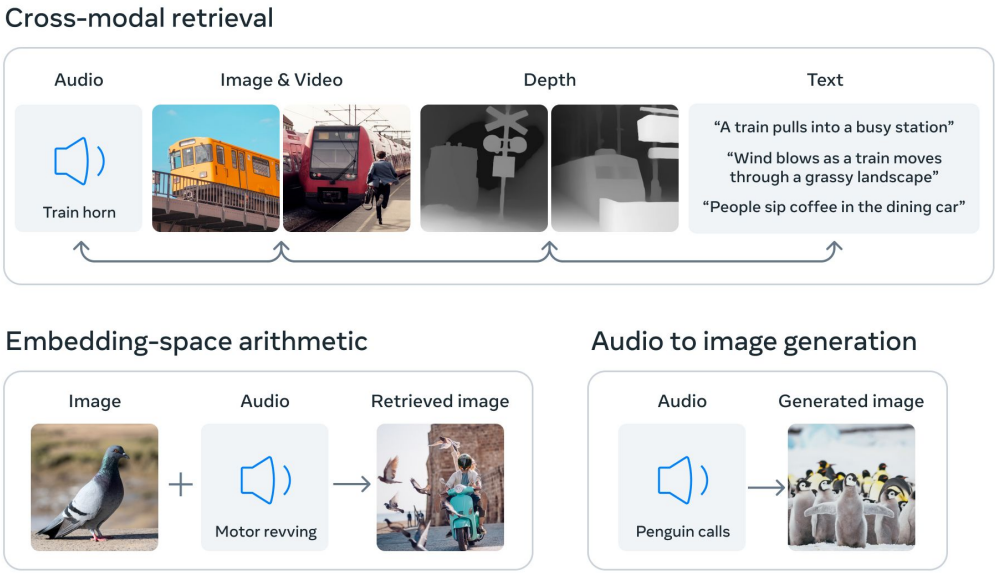


Flamingo (2022)

Source: https://arxiv.org/pdf/2204.14198

Multimodality helps LLMs learn like humans do

Larger vision models benefit non-vision tasks, such as audio classification.



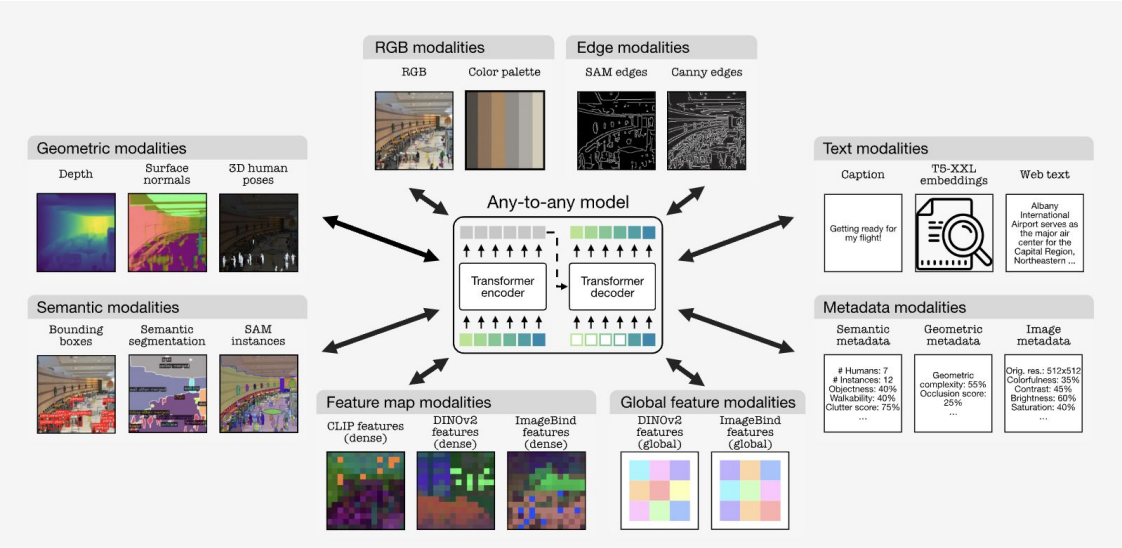
Multimodality makes LLMs smarter than mono-modality

ImageBind – binds information from six modalities

Source: https://ai.meta.com/blog/imagebind-six-modalities-binding-ai/

Deep modality – vision is more than just RGB

4M enables training a single model on tens of diverse modalities
– ***all related to images***
Joint embeddings and joint output of modalities – ***to really, really understand***

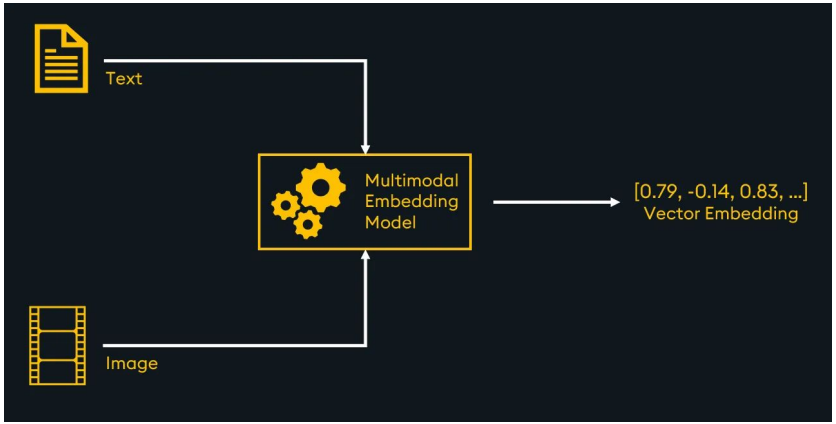


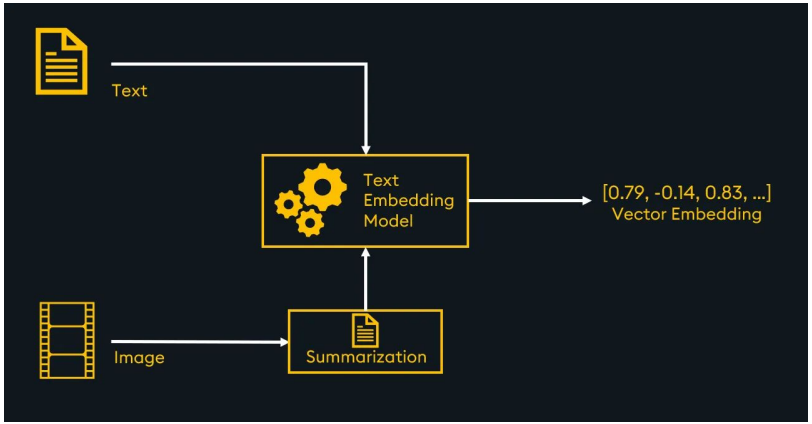
Source: https://4m.epfl.ch

How to use RAG for Multimodal Models

1. Use a multimodal embedding model to embed both text and images.

2. Use a multimodal LLM to summarize images, pass summaries and text data to a text embedding model.



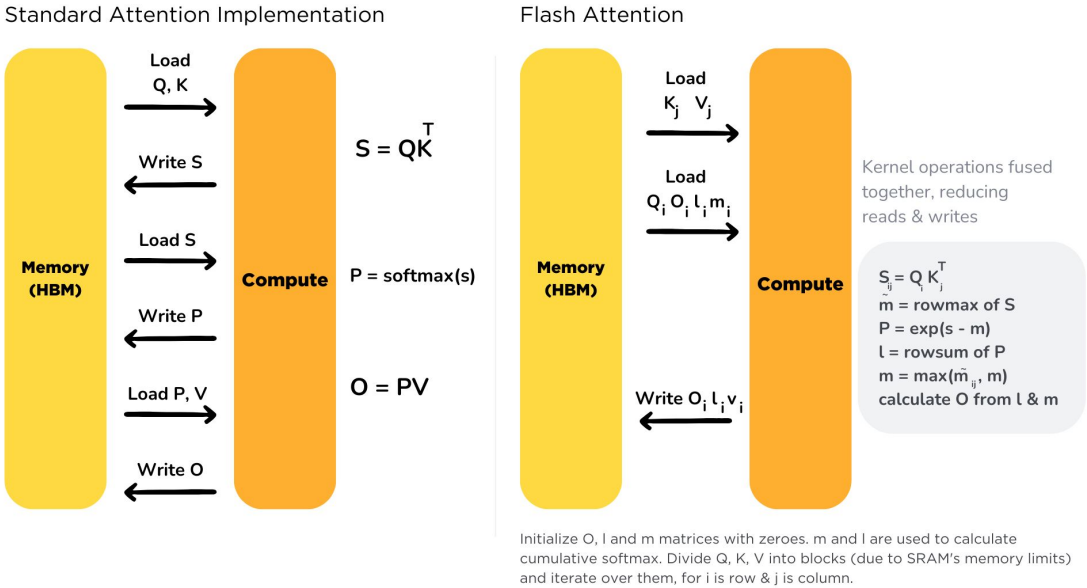


Source: https://medium.com/kx-systems/guide-to-multimodal-rag-for-images-and-text-10dab36e3117

How to speed up Transformers?

Reduce mem read-write ops

Flash Attention is a technique that **speeds up** how AI processes information by doing many steps at once instead of one at a time.

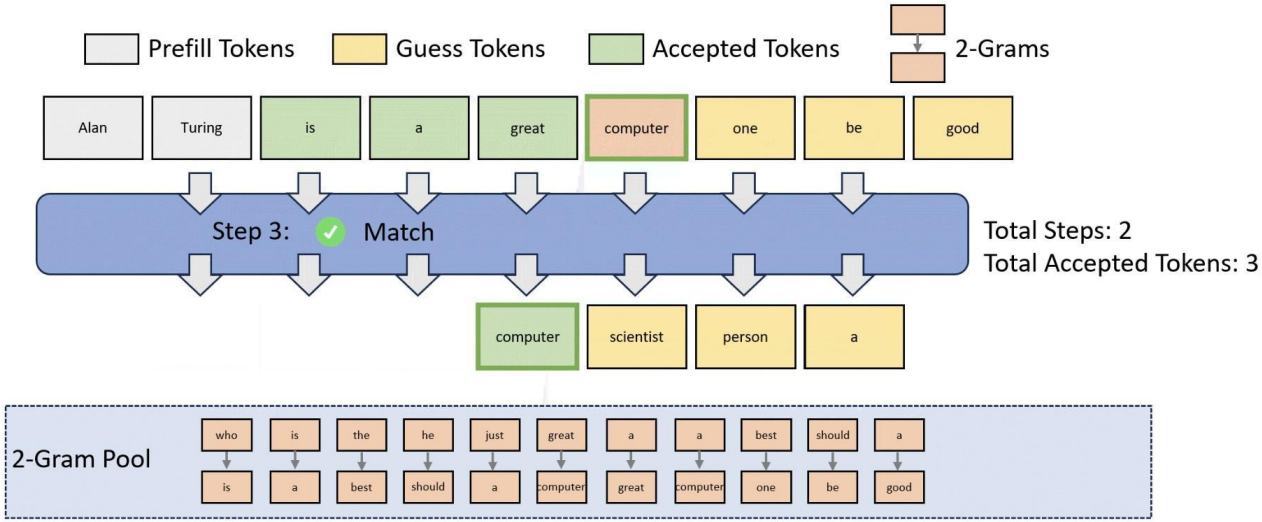


Source: https://huggingface.co/docs/text-generation-inference/en/conceptual/flash_attention

How to speed up Transformers?

Predict multiple tokens for autoregressive decoders

If we think of autoregressive decoders as solving non-linear equations, we can ***iteratively guess*** the solution, and then ***verify token matches***.

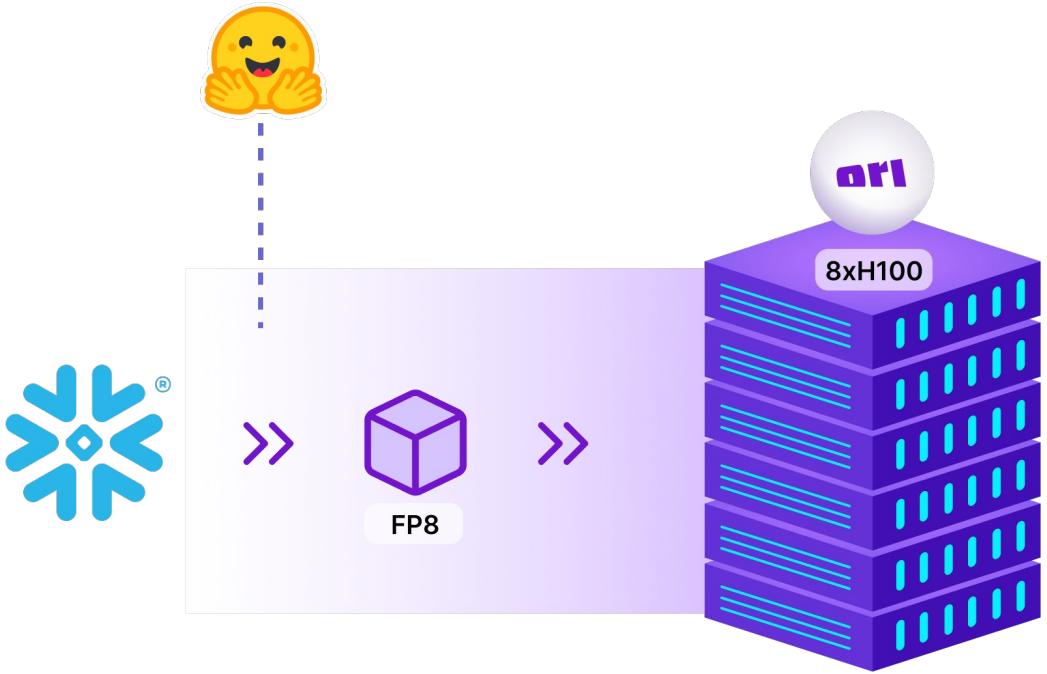


Source: https://lmsys.org/blog/2023-11-21-lookahead-decoding

Deploying a **really large** (479B) LLM

Using FP8 quantization by DeepSpeed

Running the ***Snowflake Arctic Instruct LLM*** on Ori's robust bare metal setup equipped with ***8xH100 GPUs***, for maximum efficiency and scalability.



Recommendations for your next AI Project

- ✓ Focus on Multimodal LLMs
- ✓ Even if you are using diffusion models for Image generation, consider using multimodal LLMs for editing.
- ✓ You may need to fine-tune using 4M framework
- ✓ Start collecting multimodal data that will be handy for your next application development.