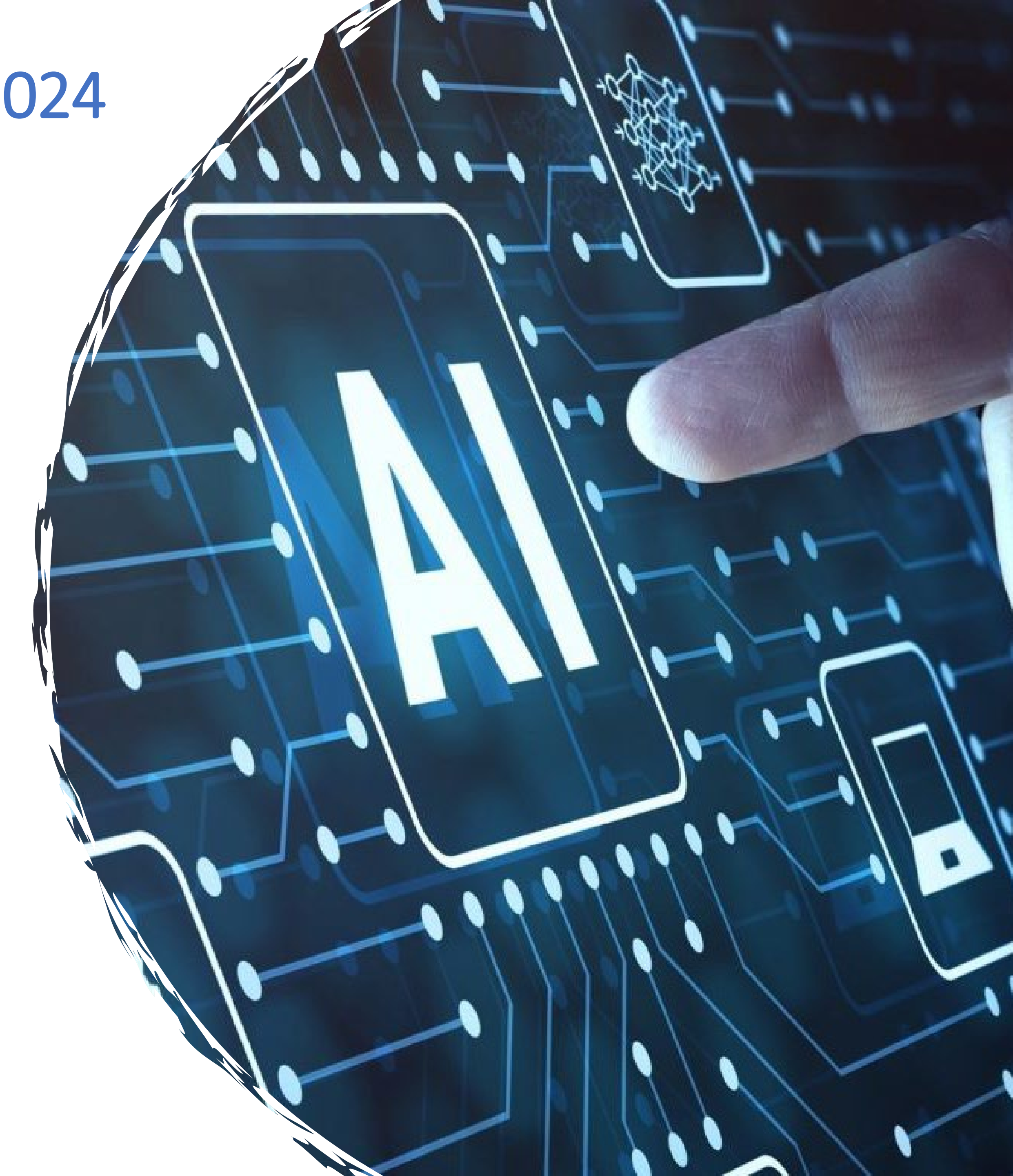
ETSI AI Conference 2024

Generative AI

From Language to Vision

Presented by: Prof. Pietro Michiardi

05/02/2024



Overview

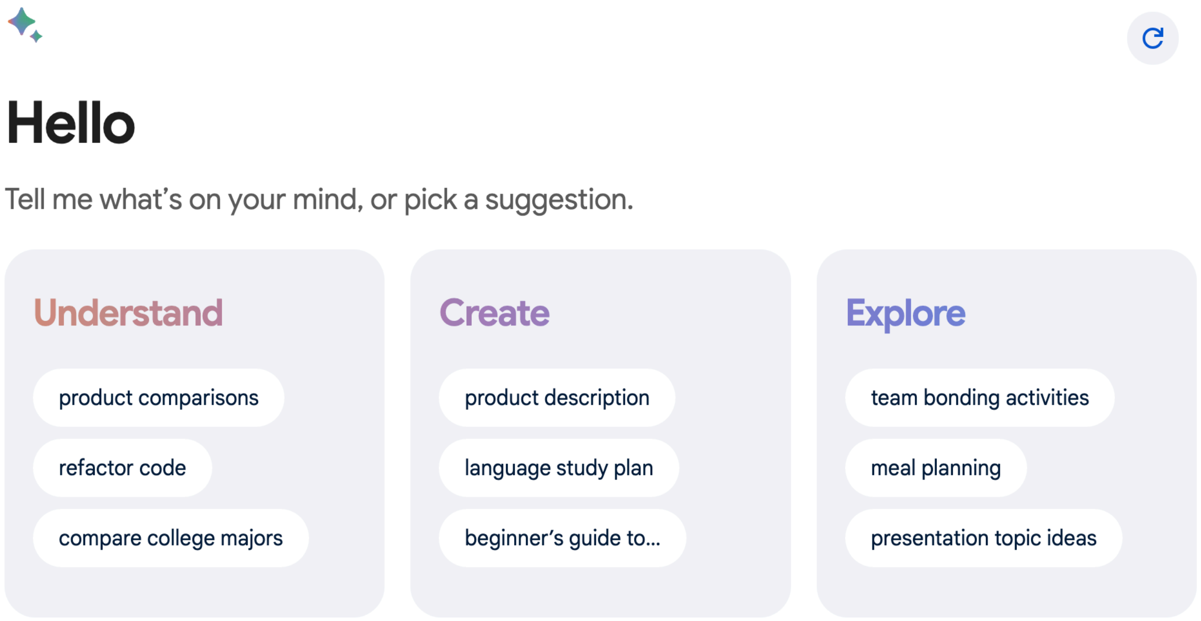
GenAI for Natural Language

- Text as sequential data
- Transformers
- LLMs, training, fine-tuning

GenAI for Computer Vision

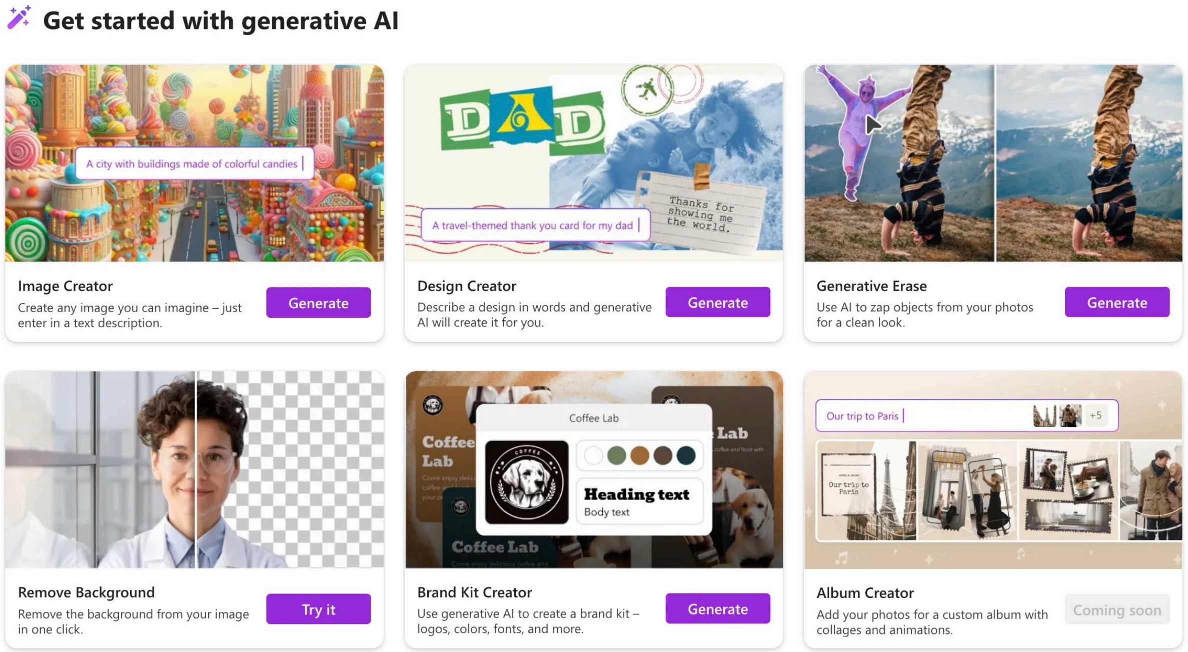
- Inspiration from physics
- Conditional generation
- Multi modality

Perspectives



Hello

Tell me what's on your mind, or pick a suggestion.



Get started with generative AI

Image Creator
Create any image you can imagine – just enter in a text description.

Design Creator
Describe a design in words and generative AI will create it for you.

Generative Erase
Use AI to zap objects from your photos for a clean look.

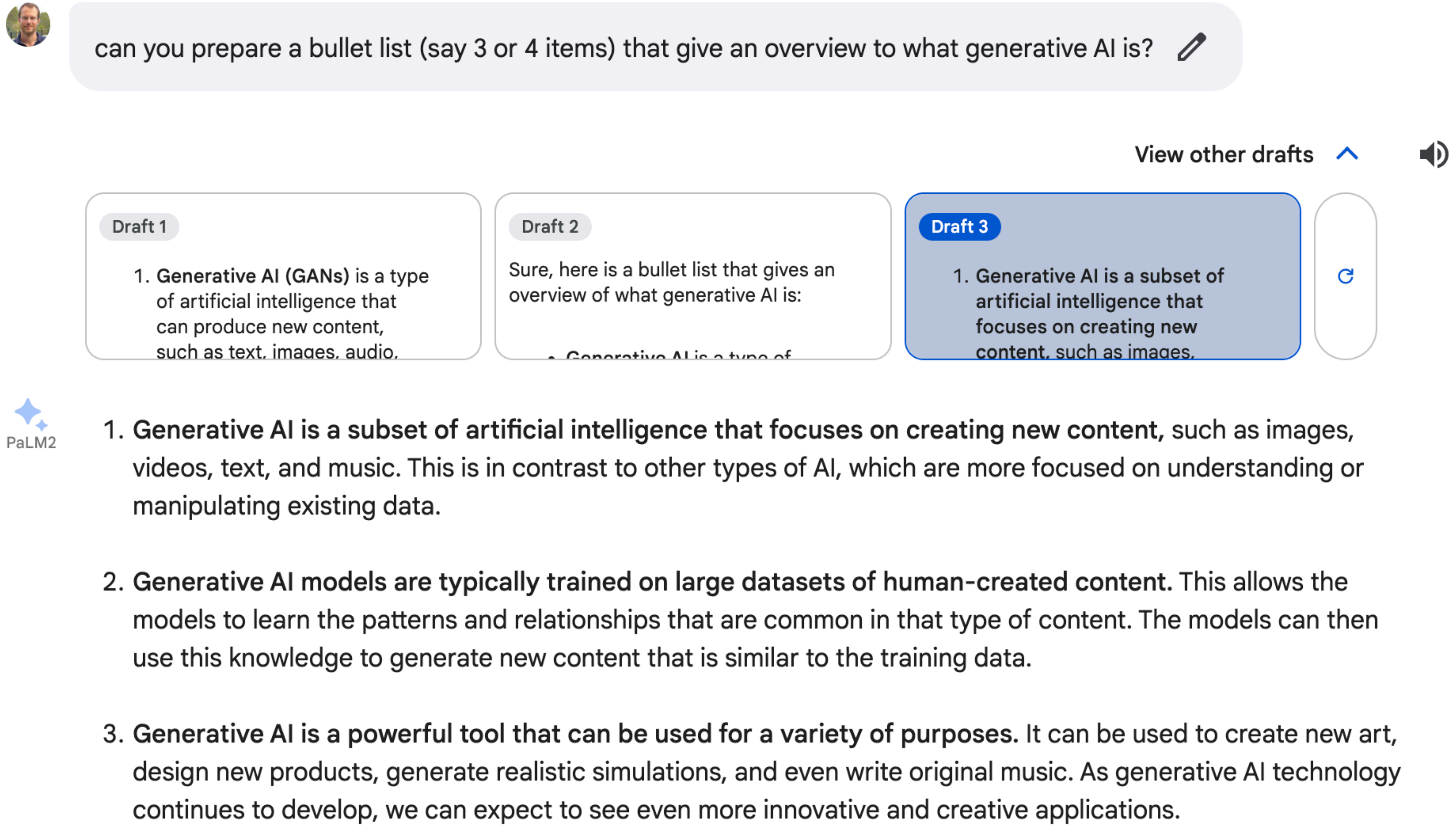
Remove Background
Remove the background from your image in one click.

Brand Kit Creator
Use generative AI to create a brand kit – logos, colors, fonts, and more.

Album Creator
Add your photos for a custom album with collages and animations.

A city with buildings made of colorful candies
Generate
DAD
A travel-themed thank you card for my dad
Thanks for showing me the world.
Generate
Generate
Try it
Coffee Lab
Heading text
Body text
Generate
Our trip to Paris
Our trip to Paris
+5
Coming soon

Generative AI: Introduction by an LLM!



can you prepare a bullet list (say 3 or 4 items) that give an overview to what generative AI is?

1. **Generative AI (GANs)** is a type of artificial intelligence that can produce new content, such as text, images, audio,

Sure, here is a bullet list that gives an overview of what generative AI is:

- Generative AI is a type of

1. Generative AI is a subset of artificial intelligence that focuses on creating new content, such as images,

1. **Generative AI is a subset of artificial intelligence that focuses on creating new content,** such as images, videos, text, and music. This is in contrast to other types of AI, which are more focused on understanding or manipulating existing data.

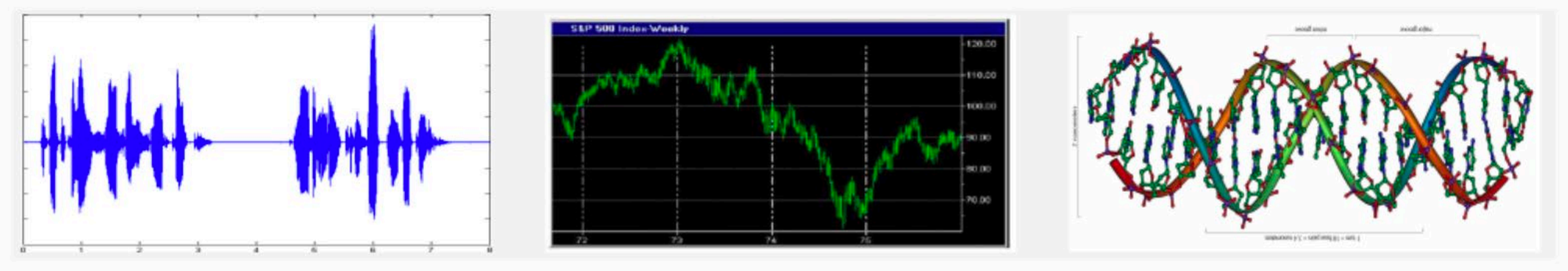
2. **Generative AI models are typically trained on large datasets of human-created content.** This allows the models to learn the patterns and relationships that are common in that type of content. The models can then use this knowledge to generate new content that is similar to the training data.

3. **Generative AI is a powerful tool that can be used for a variety of purposes.** It can be used to create new art, design new products, generate realistic simulations, and even write original music. As generative AI technology continues to develop, we can expect to see even more innovative and creative applications.

GenAI for Natural Language

Most machine learning algorithms are designed for independent data points

- Many interesting data types are not independent, for example **natural language**
- Successive points in sequential data are strongly correlated



We are interested in learning models for sequential data:

- Assume data points to be correlated
- Can use **contextual information**

What's the "context" in language? Structure, semantics, relations between words, …

Transformers Models

Super informal overview

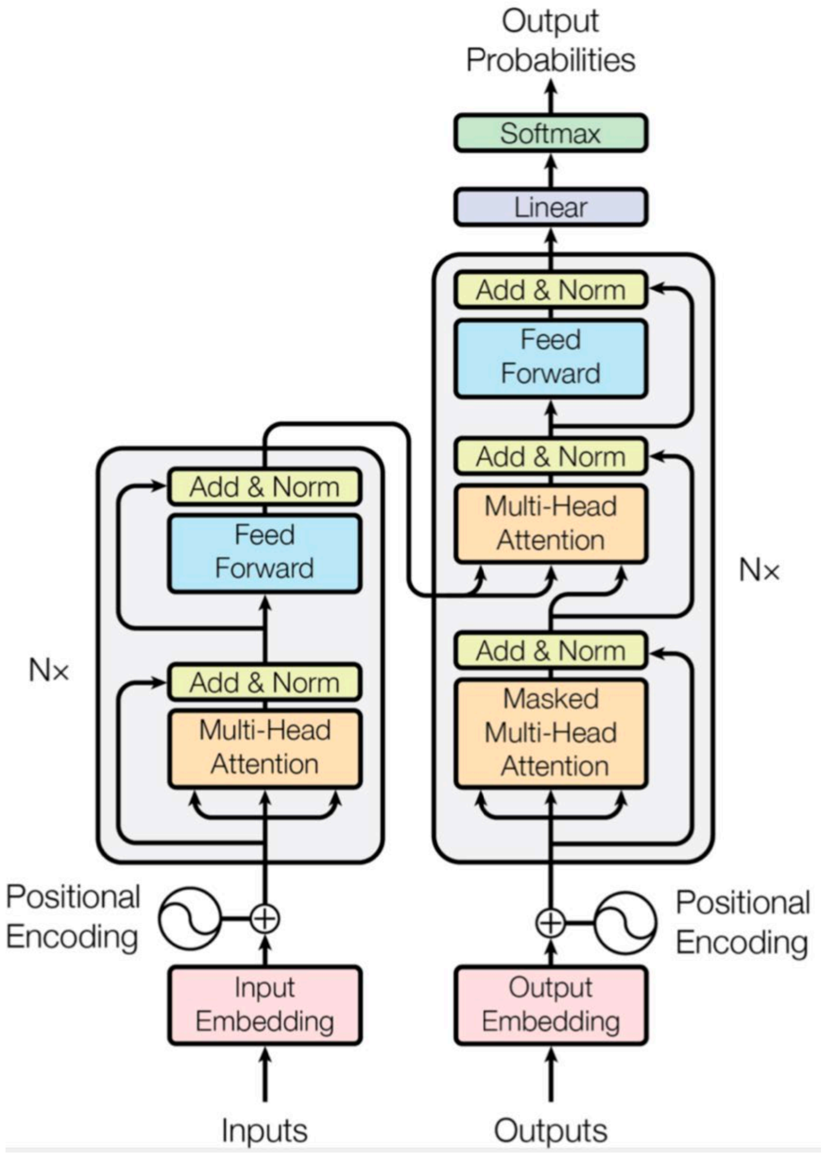
- A **paradigm shift** for sequence modeling
- No recurrence, no convolutions
- **Differentiable, probabilistic key/value store**

Advantages

- Efficient: **parallelized over thousands of GPUs**
- Easy to implement
- Flexibility in architecture design for a given task

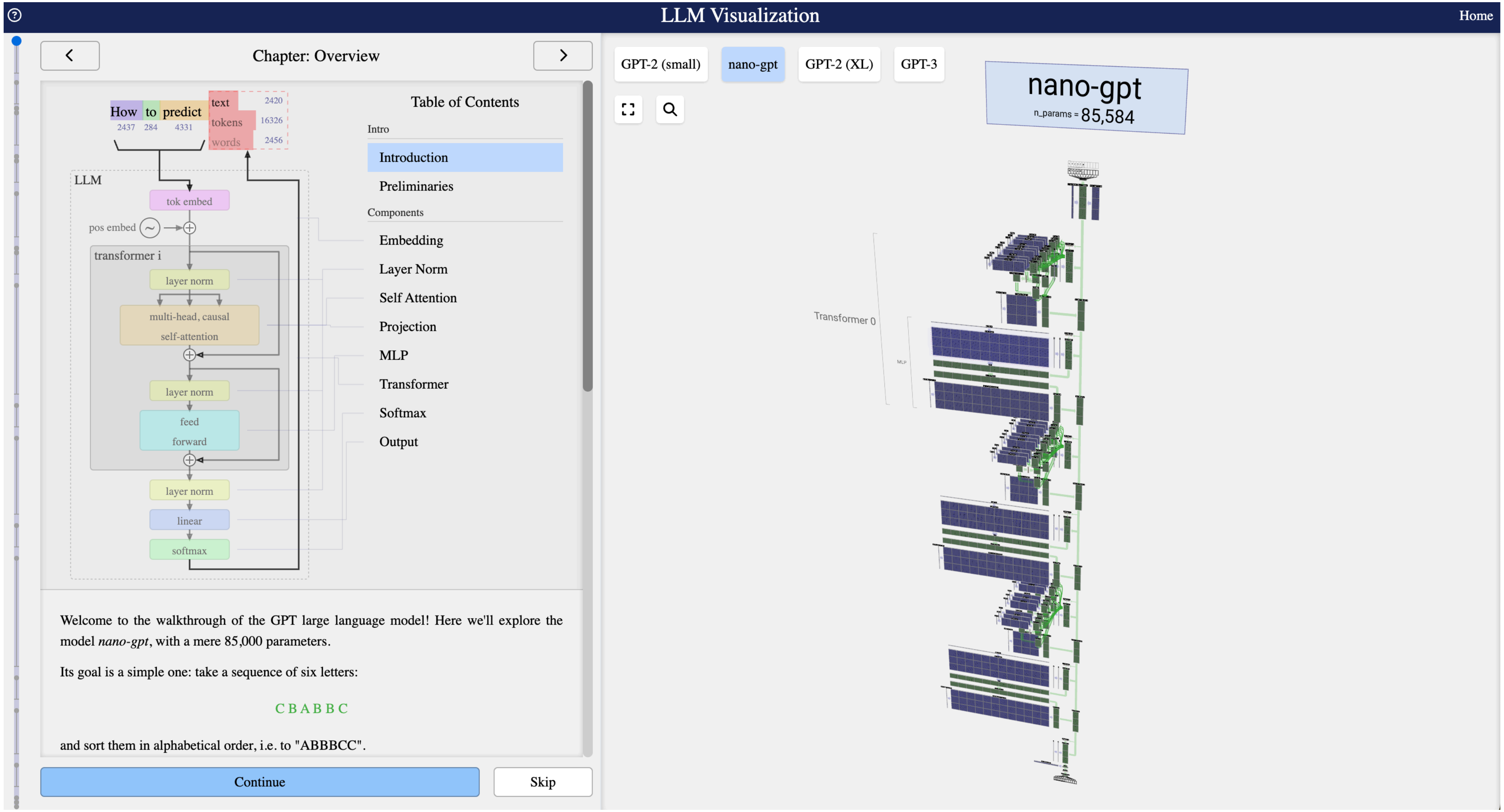
Moreover …

- All modern LLMs use (variants of) Transformers
- **Interpretable using Attention Maps**



LLM Architectures: Visualization

https://bbycroft.net/llm



Welcome to the walkthrough of the GPT large language model! Here we'll explore the model *nano-gpt*, with a mere 85,000 parameters.

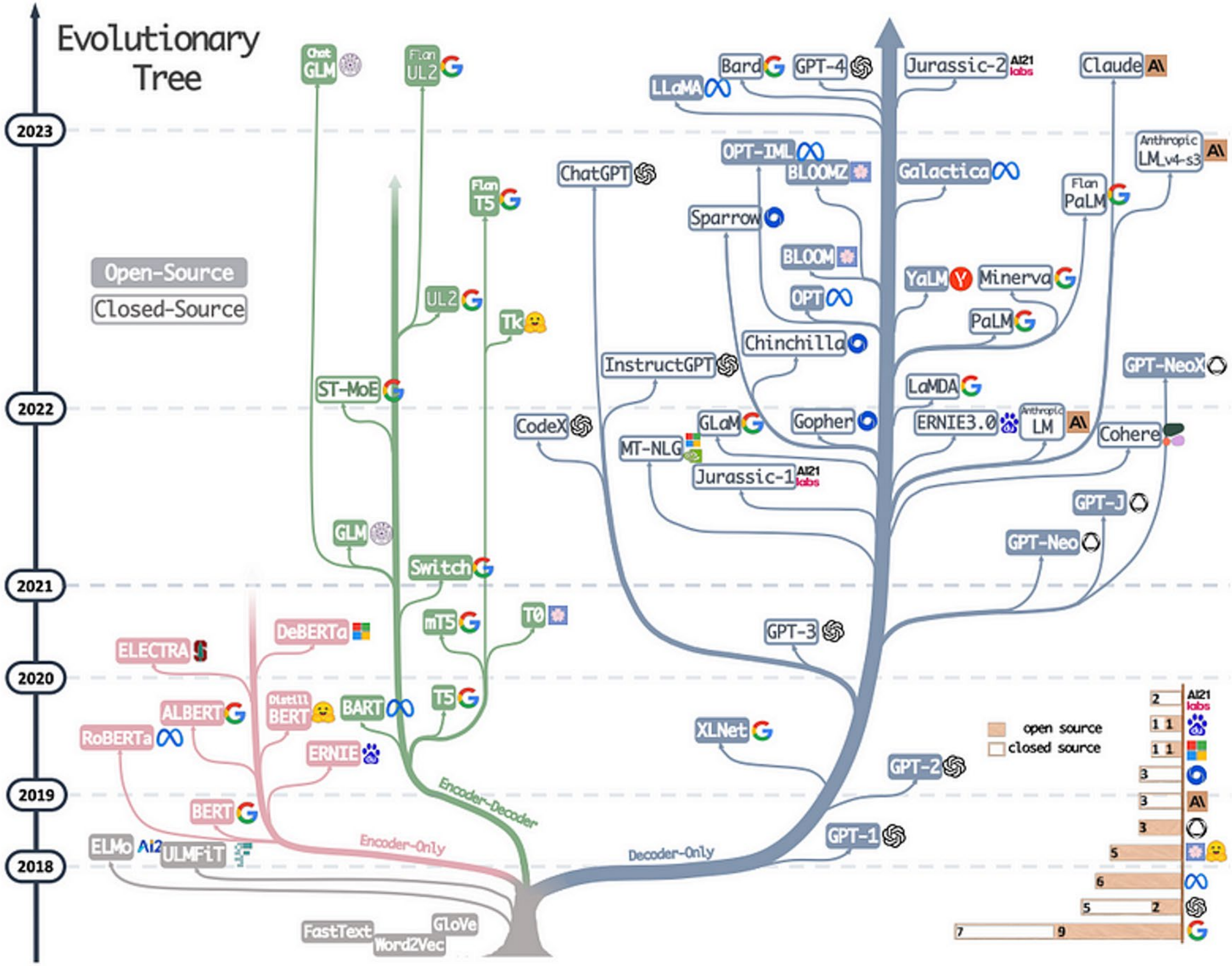
Its goal is a simple one: take a sequence of six letters:

C B A B B C

and sort them in alphabetical order, i.e. to "ABBBCC".

An explosion of open/closed source models

https://blog.sylphai.com/introduction-to-large-language-models



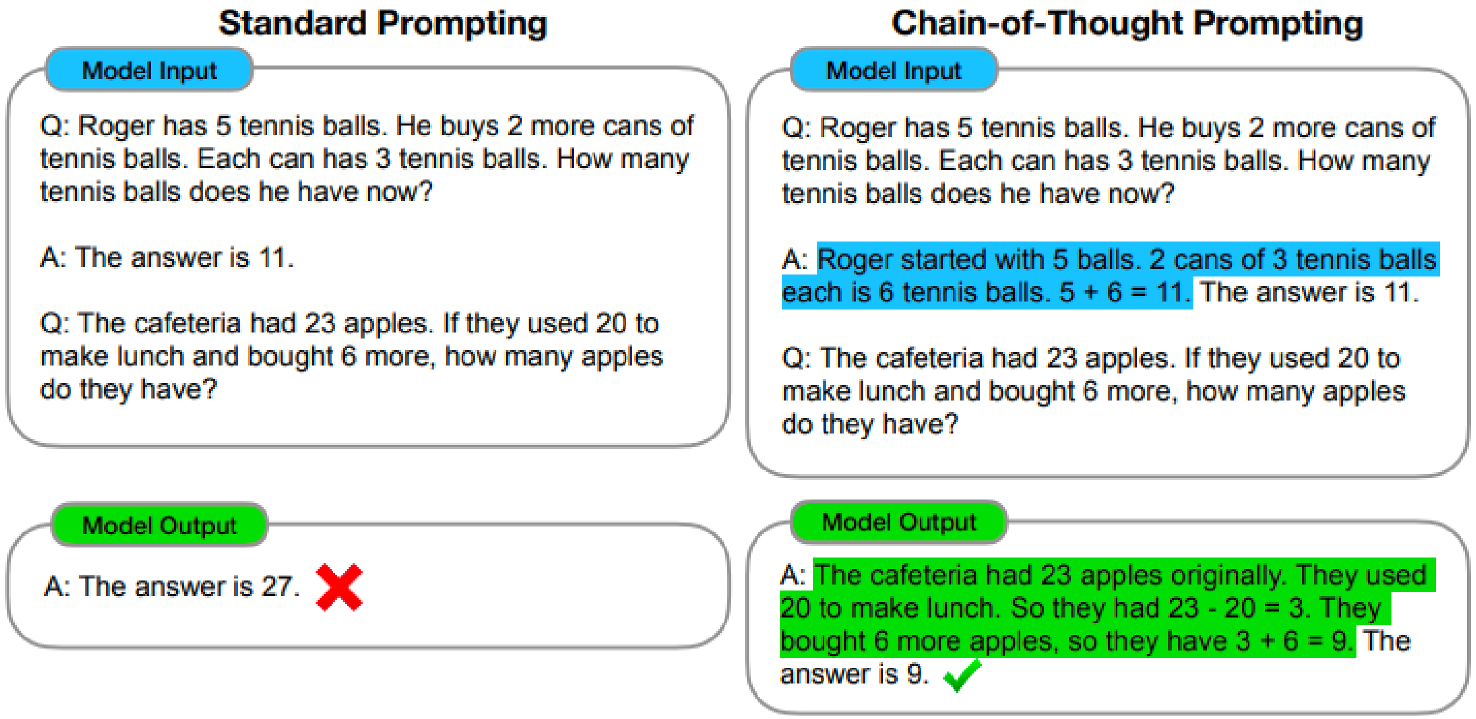
Training Strategies

General strategies

- **Autoregressive** pre-training
- **Prompts** (post-training)
  - **In-context learning**
  - Zero/Few-shot learning

Specialized strategies

- RLHF (GPT 3.5-Turbo)
- DPO and variants
- **LORA** ➔ toward fine tuning



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

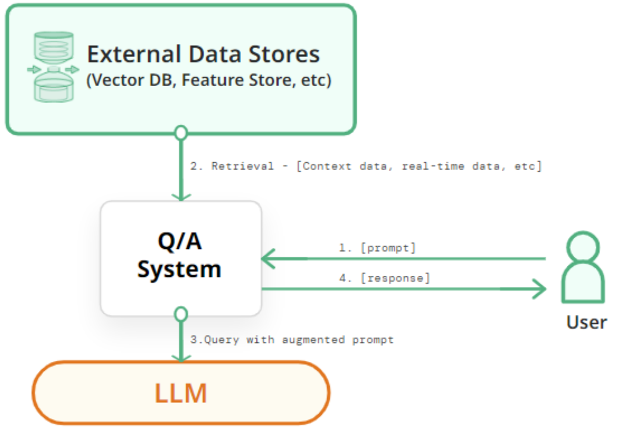
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Fine Tuning: can we use private data?


External Data Stores
(Vector DB, Feature Store, etc)
2. Retrieval - [Context data, real-time data, etc]
Q/A System
1. [prompt]
4. [response]
User
3.Query with augmented prompt
LLM

Key idea: enrich the context with indexed information

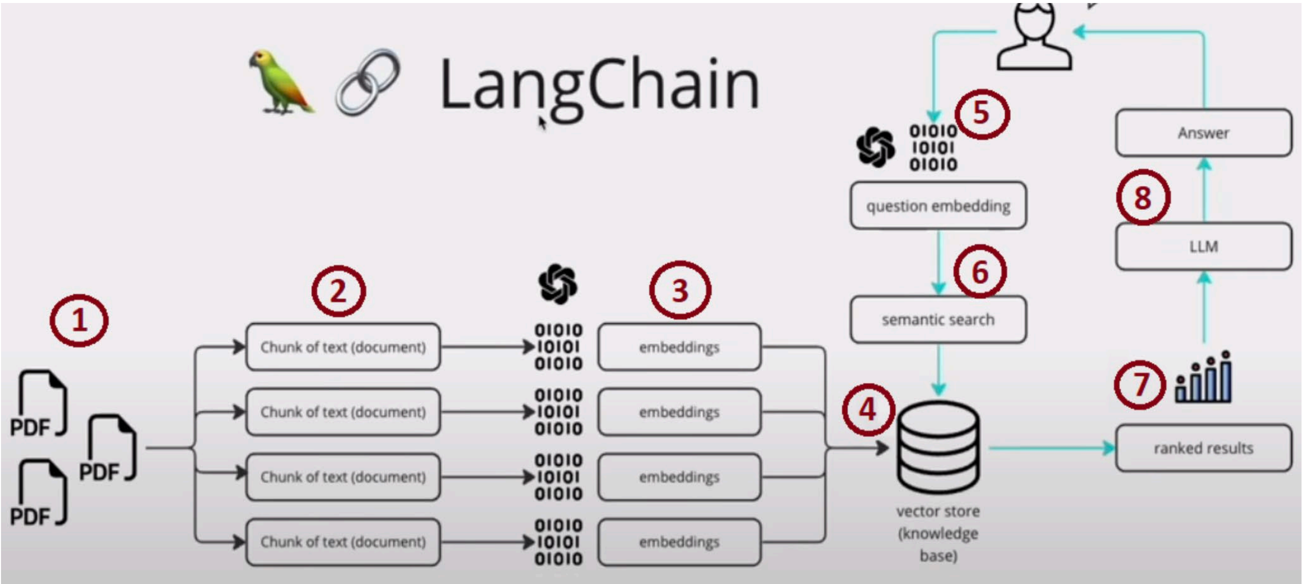
- Use your own data (many input sources!)
- Find which one is relevant given a prompt
- Inject it in the context and re-run the prompt
- Obtain **references to your own data** in LLMs answers!

LangChain, LLAMA Index, …

- Break data into chunks
- Use data and prompt **embeddings**
- Vector-DB, in-memory indexing

Careful engineering required!

- Lots of **moving parts**, **configurations**, …
- Many embedding **models to manage**



Should I own my LLM?

Computational considerations

- Be ready to become a **GPU-farm master** + **distributed systems wizard**

Data considerations

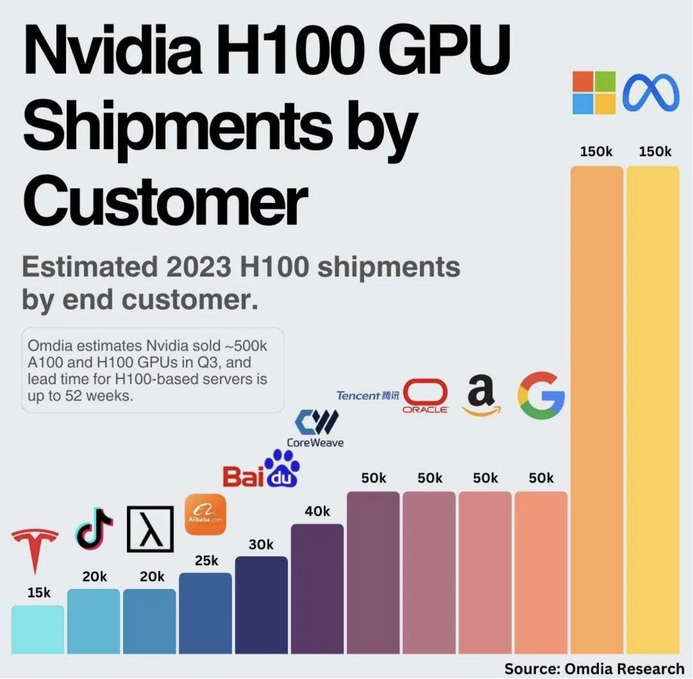
- Be ready to spend a lot of time on building **solid data pipelines**

Engineering considerations

- Highly skilled **systems engineers**, machine learning specialists

Additional considerations

- Model serving, a.k.a. inference
- **Life-cycle management**
- Common problems to deal with: **hallucination**



Nvidia H100 GPU Shipments by Customer

Estimated 2023 H100 shipments by end customer.

Omdia estimates Nvidia sold ~500k A100 and H100 GPUs in Q3, and lead time for H100-based servers is up to 52 weeks.

Source: Omdia Research

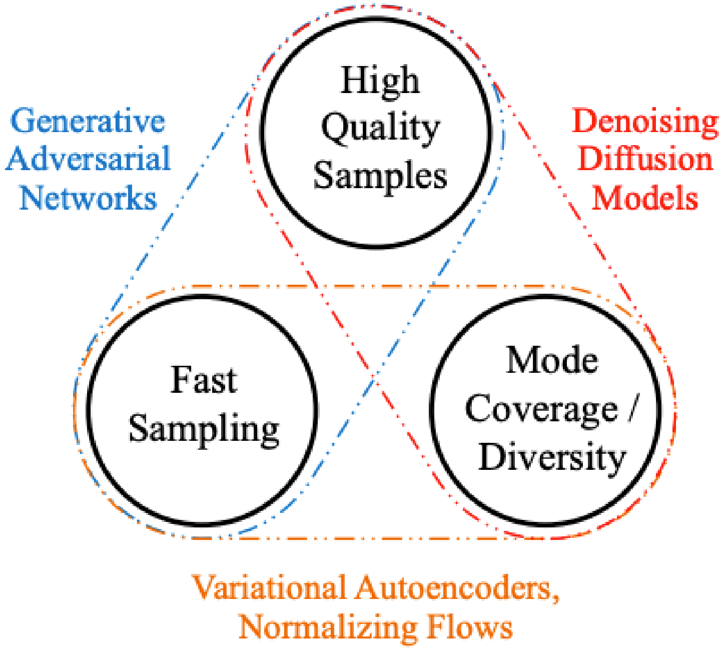
GenAI for Computer Vision

A well studied ML/CV domain

- Given new data, estimate how likely it is
- Given **noise, transform it into data**
- Examples: VAE, GANs, Normalizing Flows …

The generative trilemma

- **Quality**: often measured by **additional ML models**
- **Diversity**: difficult to assess, scoring mechanisms exist
- **Speed**: how much time to generate an image?


Generative Adversarial Networks
High Quality Samples
Denoising Diffusion Models
Fast Sampling
Mode Coverage / Diversity
Variational Autoencoders, Normalizing Flows

Diffusion Models: the Ugly

An SDE allows studying the evolution of a stochastic process $\mathbf{x}(t)$:

$$\mathbf{dx}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + \sqrt{2}\mathbf{D}(\mathbf{x}(t), t)\mathbf{dw}(t)$$

- $\mathbf{f}(\cdot, \cdot) : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$, **drift**
- $\mathbf{D}(\cdot, \cdot) : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^{N \times M}$, **diffusion matrix**
- $\mathbf{w}(t)$, **Wiener process**. Informally, $\mathbf{dw}(t) \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I})$

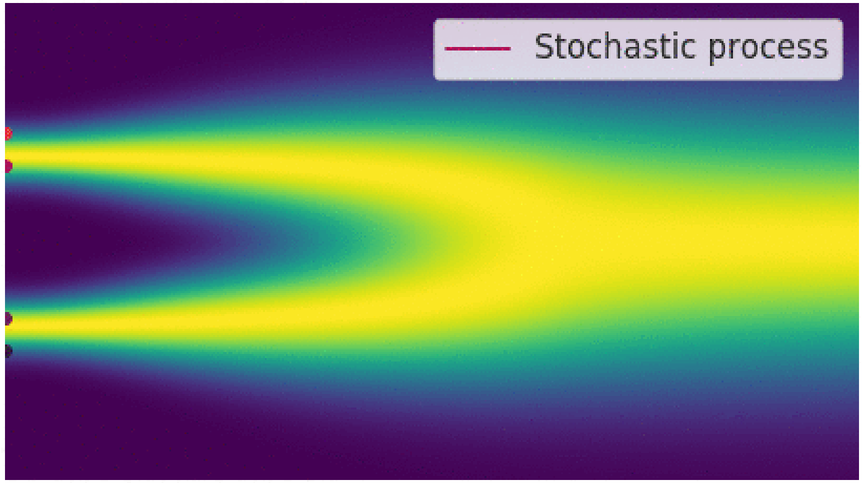
Diffusion Models: the Intuition

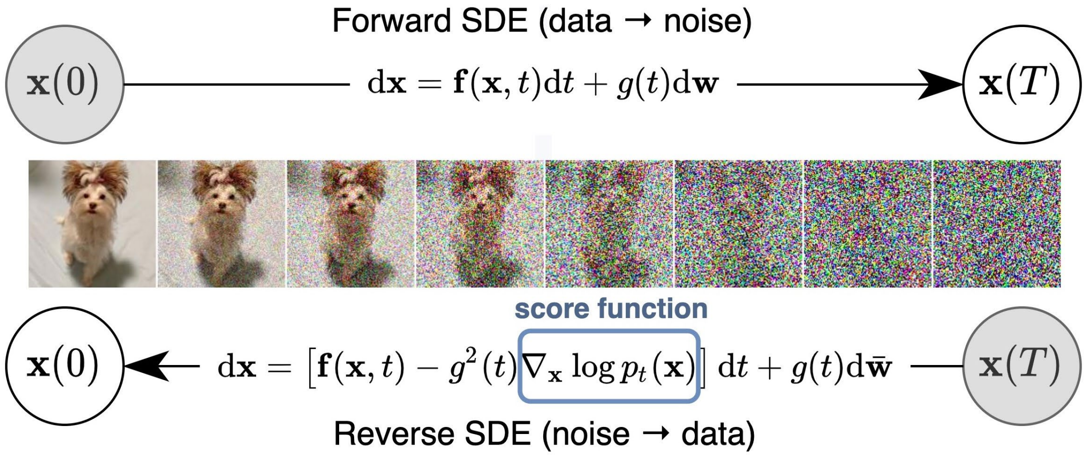
Key idea: can we use physics to generate images?

- Forward dynamics: **perturb** data **with noise**
- Backward dynamics: learn to **denoise data**

It works because we have a secret ingredient!

- Thermodynamics principles are preserved
- **The score**: once learned, we can sample noise and generate new images!





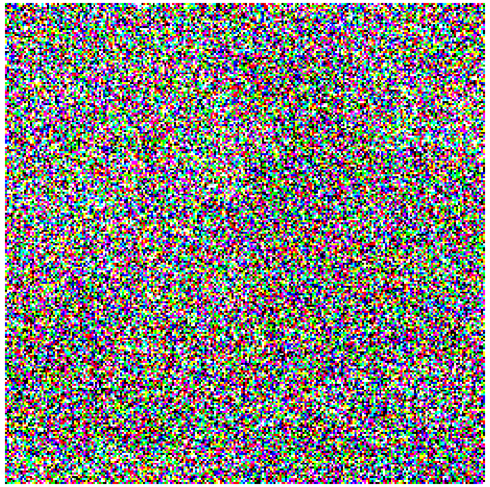
Forward SDE (data → noise)

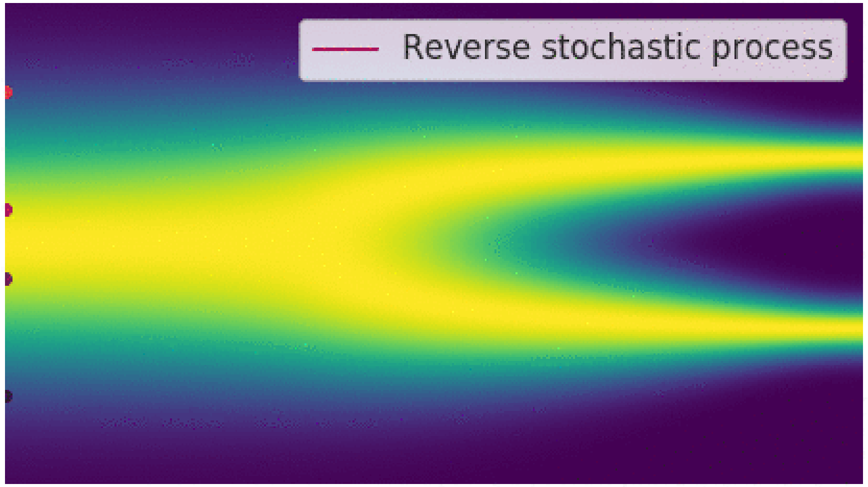
$$\mathbf{x}(0) \longrightarrow \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}(T)$$

score function

$$\mathbf{x}(0) \longleftarrow \mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}(T)$$

Reverse SDE (noise → data)





Diffusion Models: Conditional Generation

Stable diffusion

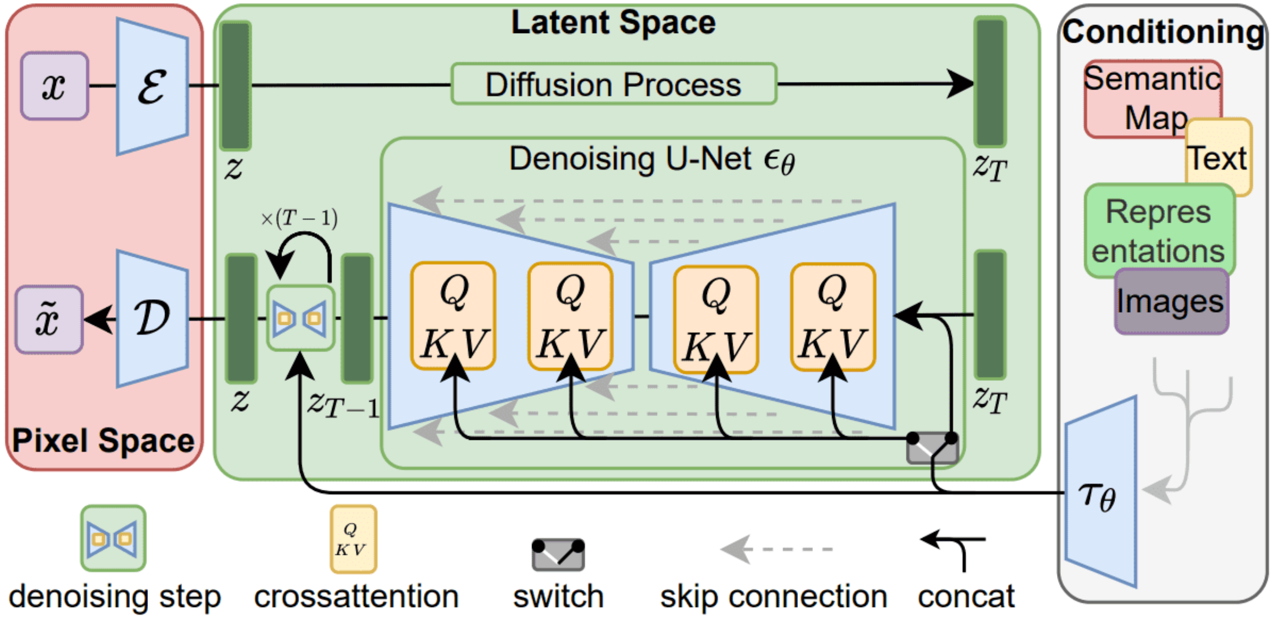
- Also known as **latent diffusion**
- Projects data to a latent space
- Denoiser + **attention instills semantics**

Conditional generation

- Allows incorporating external signal
- E.g.: text prompt **steers image generation**

Prompt engineering

- More of an art than science
- Consistent results are not easy to obtain!
- New trend: (live) **image editing**



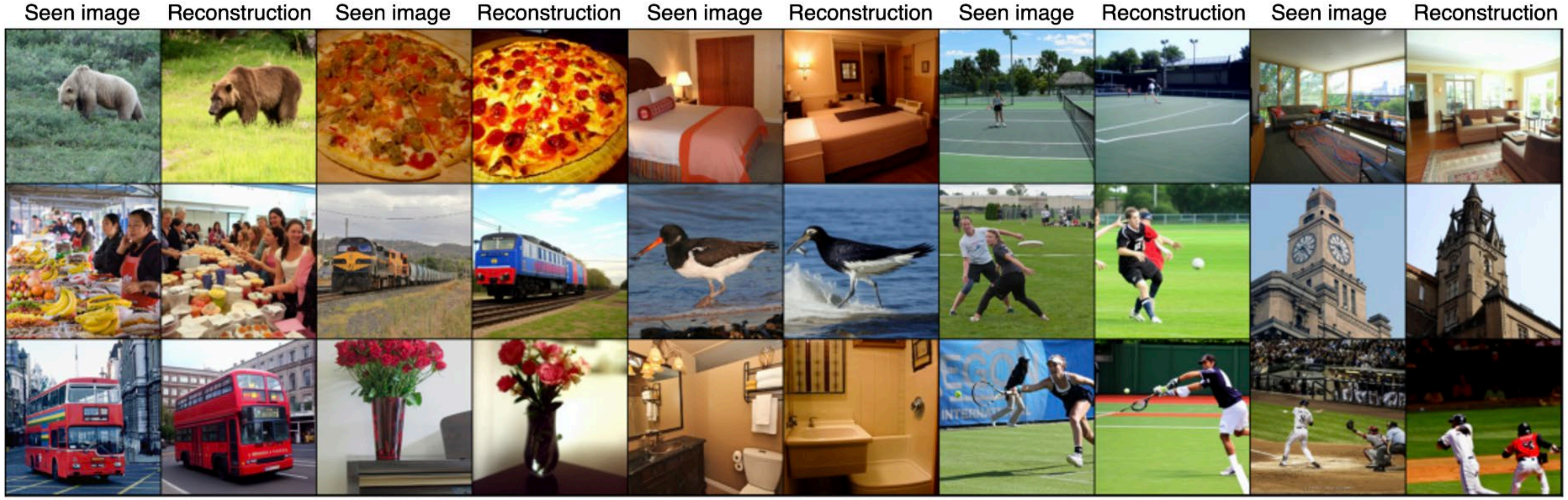
Diffusion Models: Unusual Applications!

Reconstructing the Mind’s Eye:

fMRI-to-Image with Contrastive Learning and Diffusion Priors

Paul S. Scotti[1,2*], Atmadeep Banerjee[2*], Jimmie Goode[†2], Stepan Shabalin[2], Alex Nguyen[1], Ethan Cohen[3], Aidan J. Dempster[4], Nathalie Verlinde[1], Elad Yundler[5], David Weisberg[1,2], Kenneth A. Norman[‡1], and Tanishq Mathew Abraham[‡2,6,7]

[1]Princeton Neuroscience Institute, [2]Medical AI Research Center (MedARC), [3]Ecole Normale Supérieure, PSL University, [4]University of Toronto, [5]Hebrew University of Jerusalem, [6]EleutherAI, [7]Stability AI



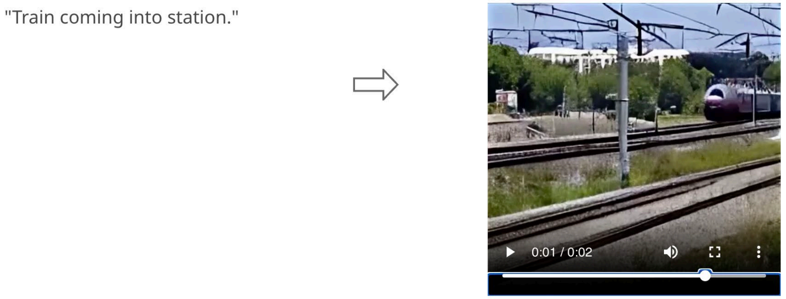
Multimodal Diffusion Models

Audio + Image → Text + Image

"Playing piano in a forest."

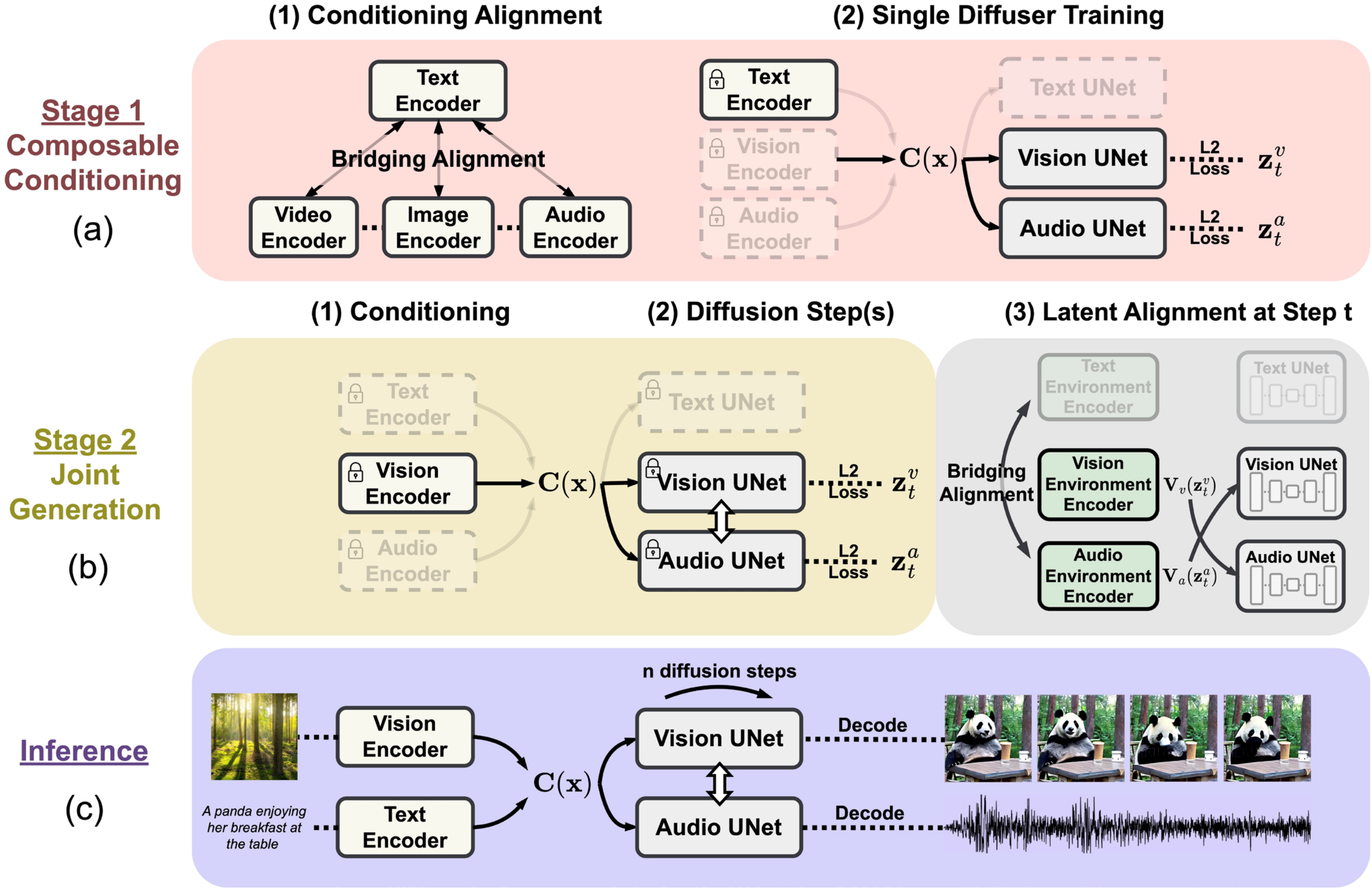
Text → Video + Audio

"Train coming into station."



Text + Audio → Video

"Forward moving camera view."



Stage 1
Composable Conditioning
(a)

(1) Conditioning Alignment

Text Encoder
Bridging Alignment
Video Encoder
Image Encoder
Audio Encoder

(2) Single Diffuser Training

Text Encoder
Vision Encoder
Audio Encoder
$C(\mathbf{x})$
Text UNet
Vision UNet
Audio UNet
L2 Loss $\mathbf{z}_t^v$
L2 Loss $\mathbf{z}_t^a$

Stage 2
Joint Generation
(b)

(1) Conditioning

Text Encoder
Vision Encoder
Audio Encoder

(2) Diffusion Step(s)

$C(\mathbf{x})$
Text UNet
Vision UNet
Audio UNet
L2 Loss $\mathbf{z}_t^v$
L2 Loss $\mathbf{z}_t^a$

(3) Latent Alignment at Step t

Bridging Alignment
Text Environment Encoder
Vision Environment Encoder
Audio Environment Encoder
$V_v(\mathbf{z}_t^v)$
$V_a(\mathbf{z}_t^a)$
Text UNet
Vision UNet
Audio UNet

Inference
(c)

Vision Encoder
A panda enjoying her breakfast at the table
Text Encoder
$C(\mathbf{x})$
n diffusion steps
Vision UNet
Audio UNet
Decode
Decode

Conclusion

Exciting new technology, some claim major industrial revolution

- **Integration** with MS Office, Google Workspace, etc …
- GenAI for source code generation, analysis, bug fixing, etc …
- **Augmented intelligence**! Not there to "replace" humans!

Technology is still in its infancy

- **Causal reasoning**, world model are not there yet
- Use tons of human-generated data. **Memorization? Legal issues?**

Way ahead …

- Quantification of **uncertainty: your model should know when it does not know**
- Better alignment of prompts to images: **Generative Semantic Nursing**
- **Multi-modality** for conversational agents

Thank you!

Prof. Pietro Michiardi

EURECOM
Sophia Antipolis