Deep Generative Models

Generative AI Systems

Hamid Beigy

Sharif University of Technology

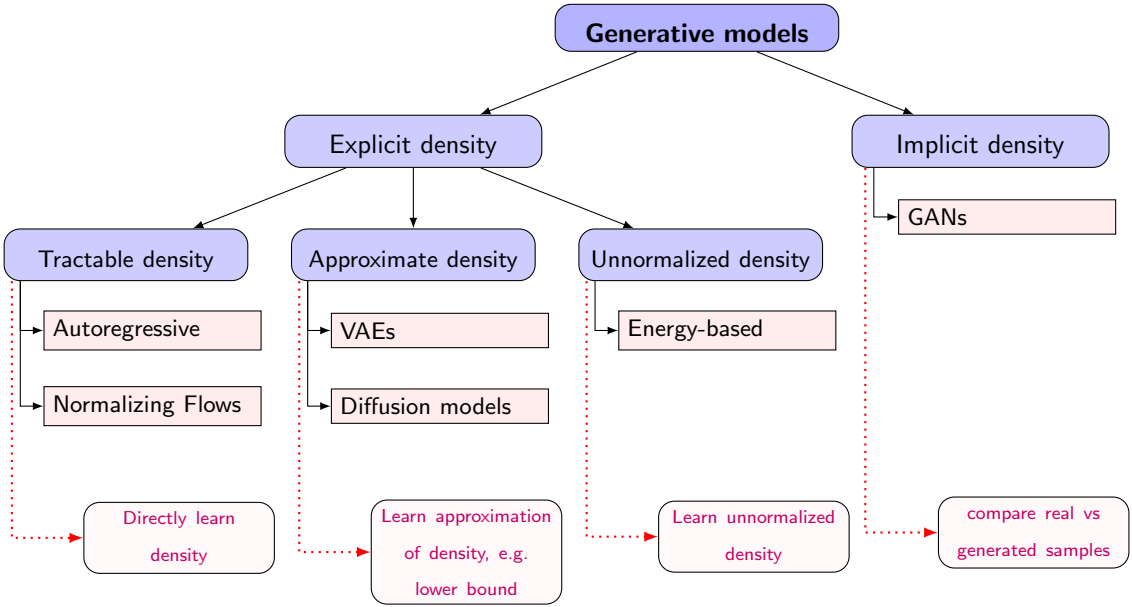
May 30, 2025

Table of contents

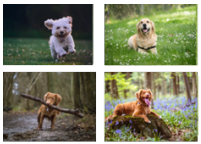
1. Introduction
2. Large Language models
3. Multi-modal Generative models
4. Generative AI Systems
5. Summary
6. References

Introduction

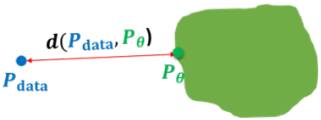
Generative models categories



1. Assume that the observed variable $\mathbf{x}$ is a random sample from an underlying process, whose true distribution $p_{data}(\mathbf{x})$ is unknown.



$\mathbf{x}_i \sim P_{\text{data}}$
$i = 1, 2, \ldots, n$



2. We attempt to approximate this process with a chosen model, $p_\theta(\mathbf{x})$, with parameters θ such that $\mathbf{x} \sim p_\theta(\mathbf{x})$.

3. Learning is the process of searching for the parameter θ such that $p_\theta(\mathbf{x})$ well approximates $p_{data}(\mathbf{x})$ for any observed $\mathbf{x}$, i.e.

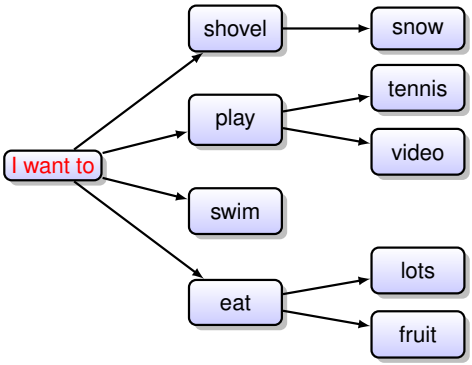
$$p_\theta(\mathbf{x}) \approx p_{data}(\mathbf{x})$$

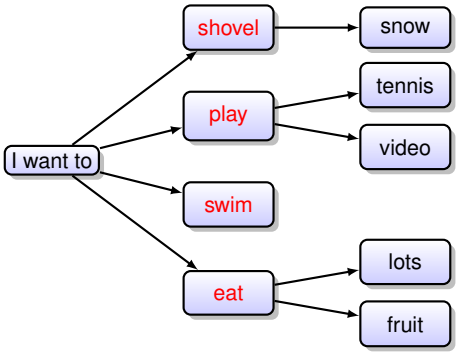
4. We wish $p_\theta(\mathbf{x})$ to be sufficiently flexible to be able to adapt to the data for obtaining sufficiently accurate model and to be able to incorporate prior knowledge.

Credit: Aditya Grover

Large Language models

A language model is a model for how humans generate language.





The **language modeling** task is:
Given sequence of words so far (**context**), **predict what comes next**.

Transformers model

1. The attention make it possible to do sequence to sequence modeling without recurrent network units (Vaswani et al. 2017).
2. The transformer model is entirely built on the self-attention mechanisms without using sequence-aligned recurrent architecture.

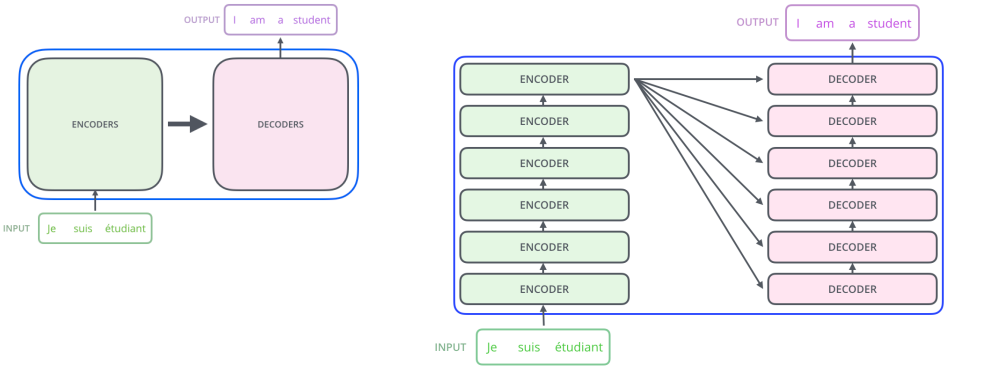


Figure: Jay Alammar

3. The encoding component is a stack of six encoders.
4. The decoding component is a stack of decoders of the same number.

Transformers training

1. The Transformers works slightly differently during training and inference.
2. Input sequence: You are welcome in English.
3. Target sequence: De nada in Spanish

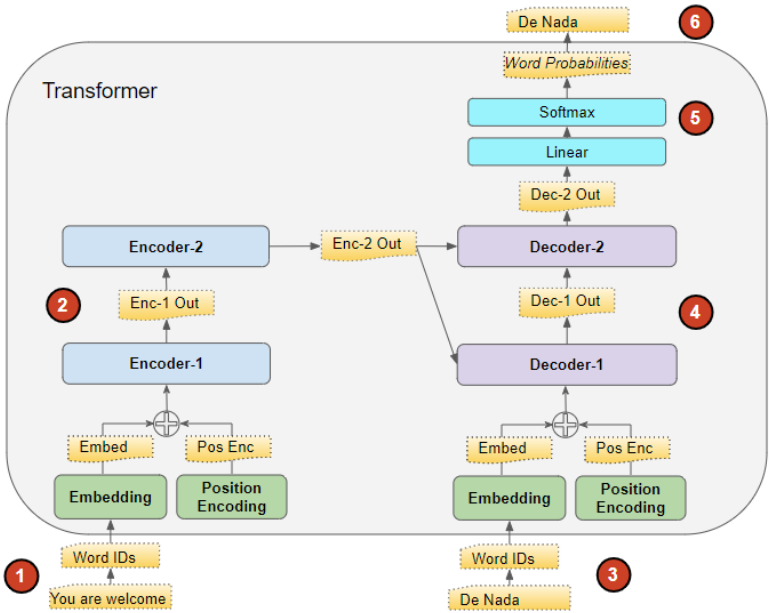


Figure:Ketan Doshi

Transformers inference

1. During Inference, we have only the input sequence and don't have the target sequence to pass as input to the Decoder.
2. The goal is to produce the target sequence from the input sequence alone.

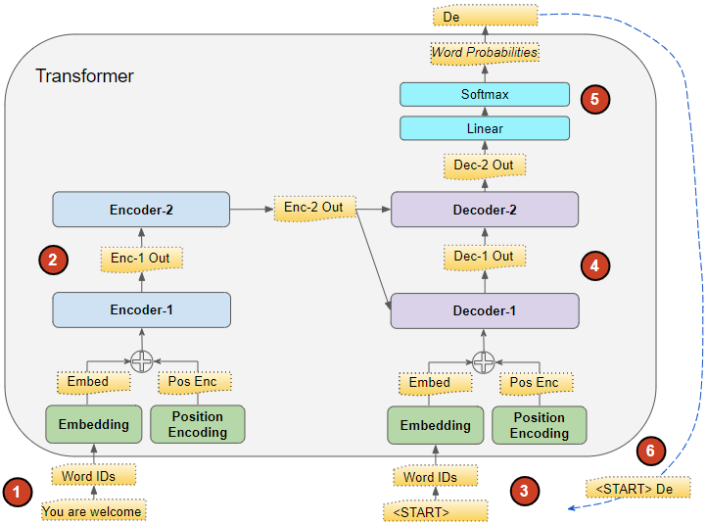
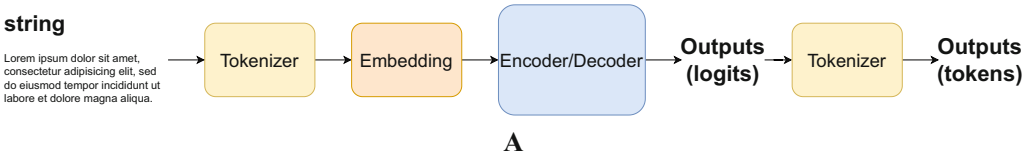
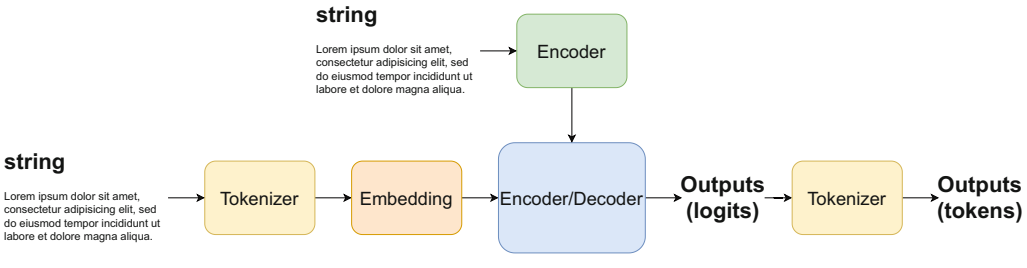


Figure:Ketan Doshi

General Architectures of LLMs



A

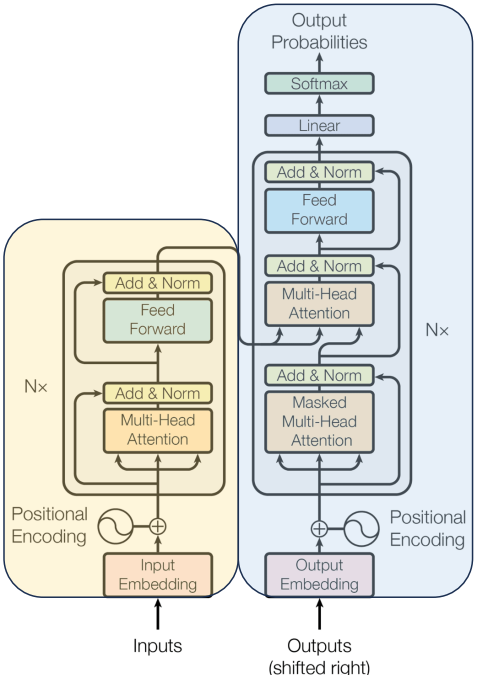


B

1. Language models can be used to perform multiple tasks by learning to predict the next token or sentence

BERT
Oct 2018

Representation



GPT
Jun 2018

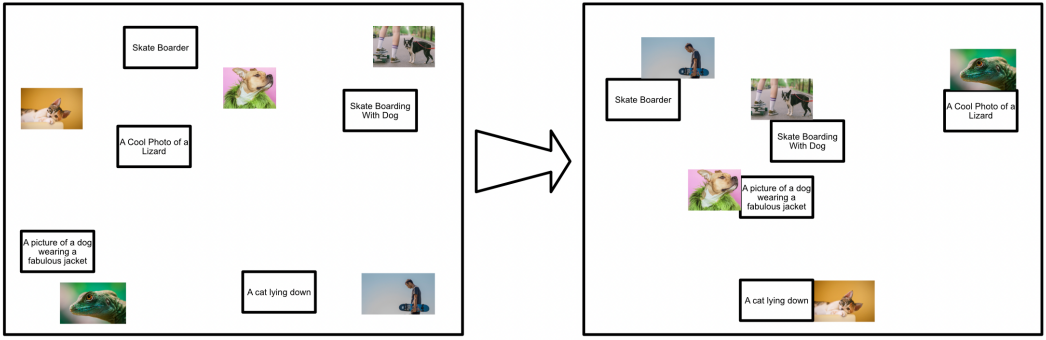
Generation

Figure:Roshan Sharmai

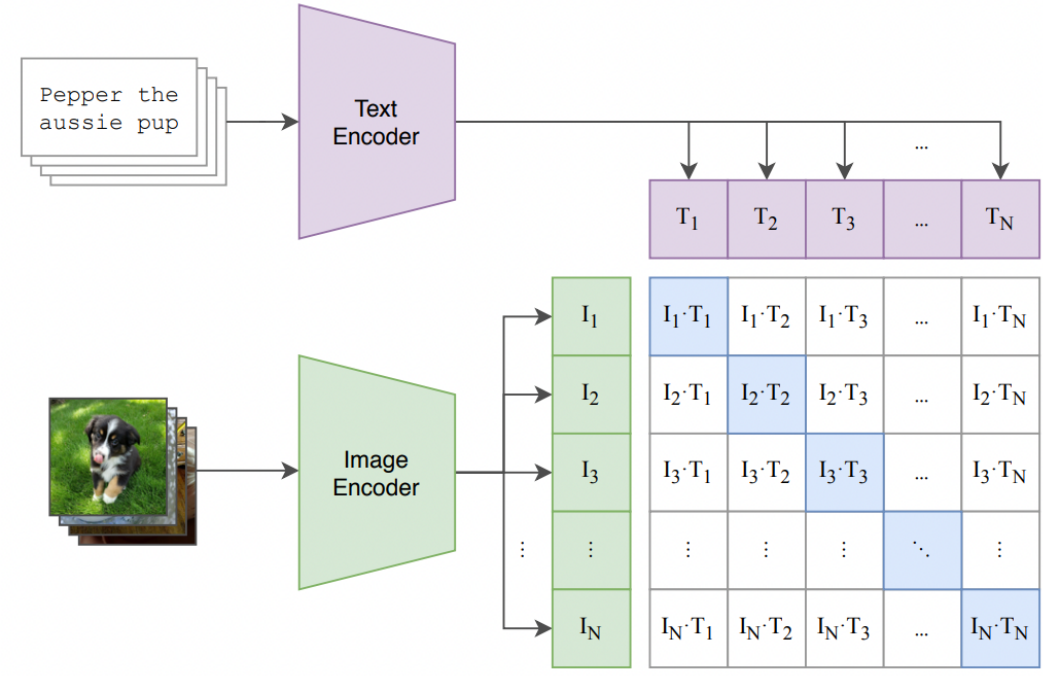
Multi-modal Generative models

CLIP model

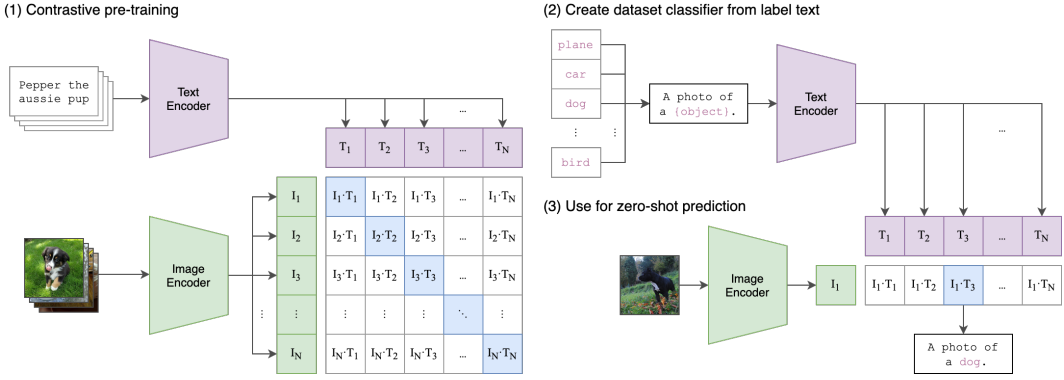
1. CLIP stands for contrastive language-image pre-training.
2. The core idea of CLIP is to use captioned images scraped from the Internet to create a model which can predict if text is compatible with an image or not.



In CLIP uses contrastive learning to learn a text encoder and an image encoder.

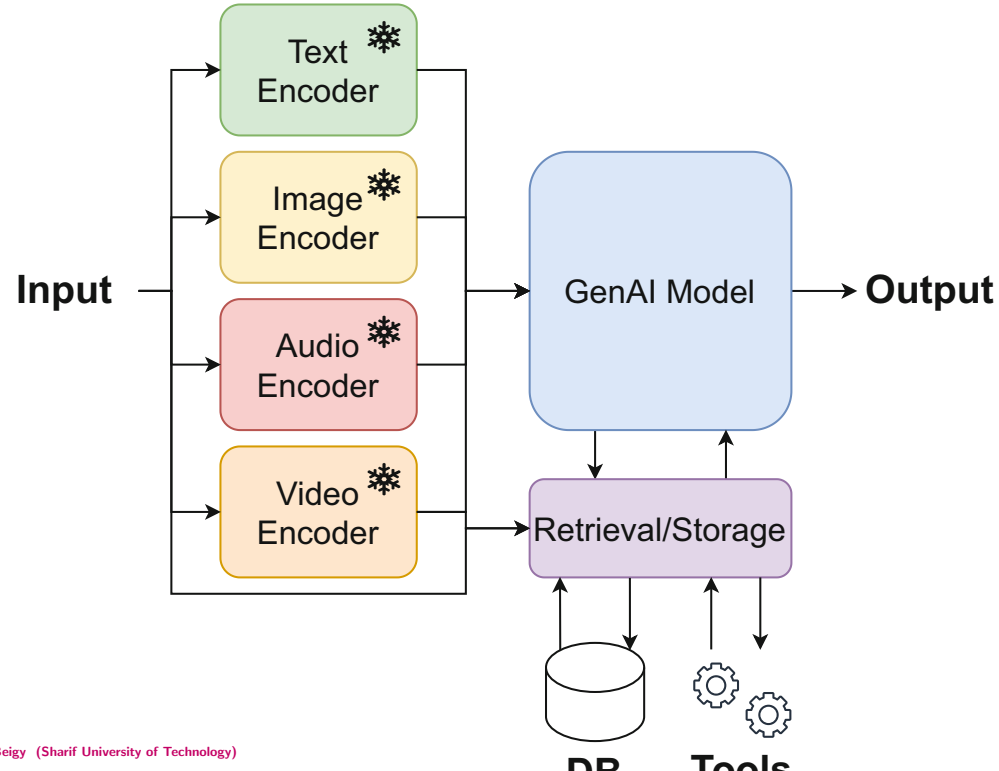


CLIP model


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

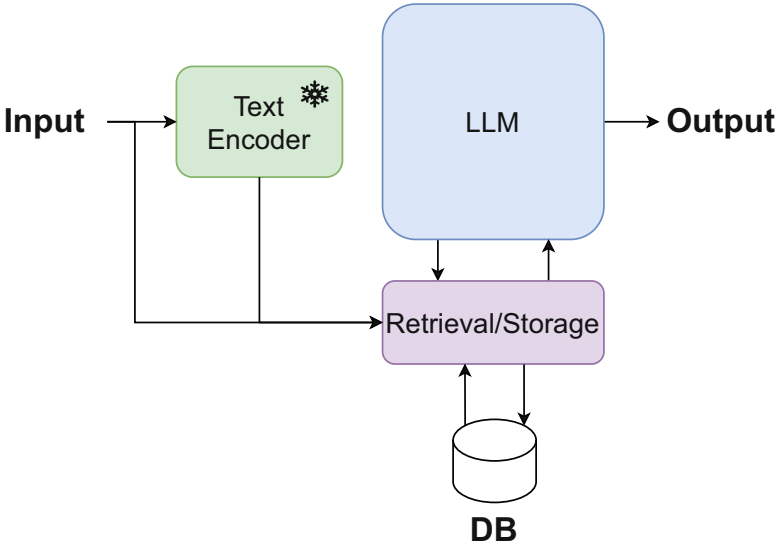
Generative AI Systems

General Architecture

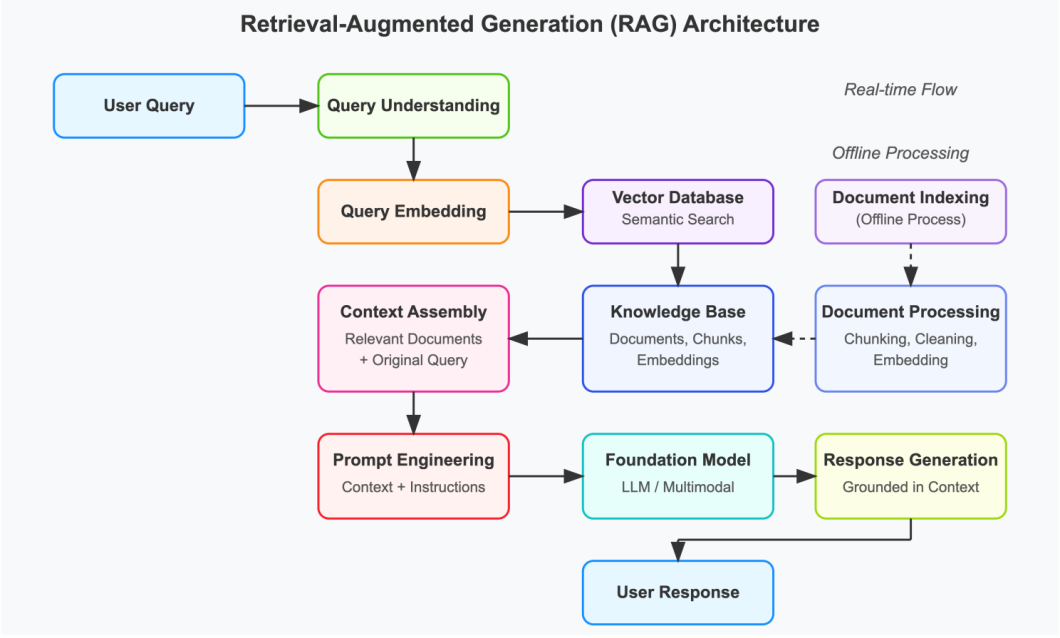


Example: Retrieval Augmented Generations

1. The idea is based on utilizing a database of texts and two LLMs (Lewis et al. 2020):
   - an encoder-LLM
   - a decoder-LLM

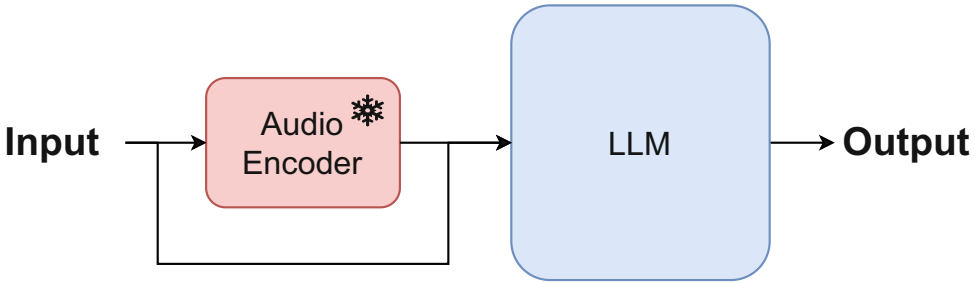


Example: Retrieval Augmented Generations



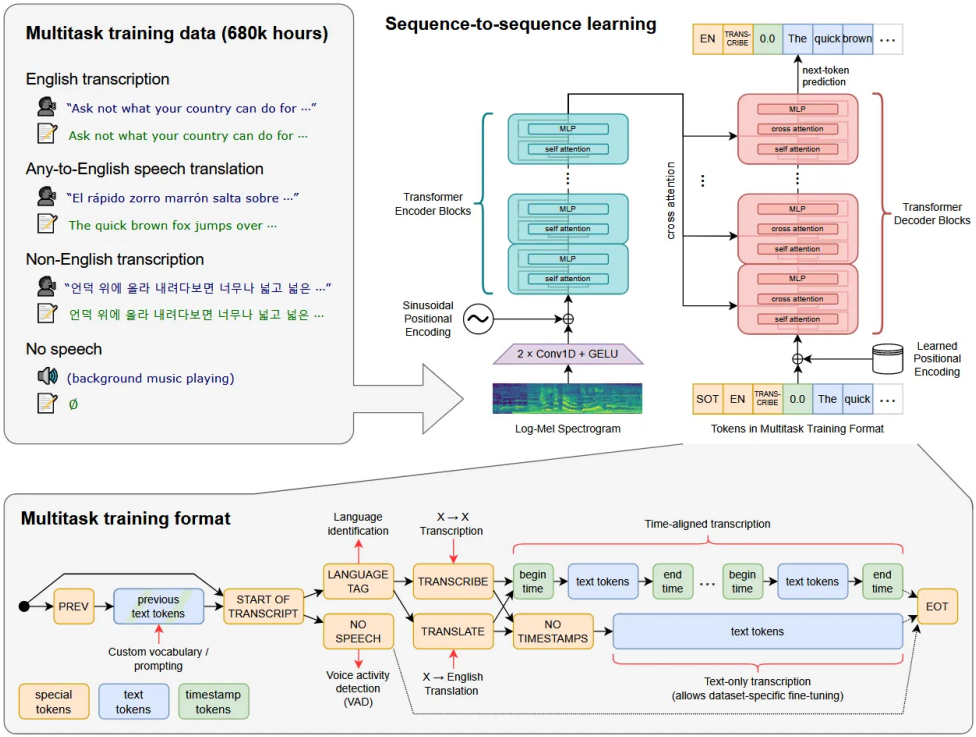
Example: Speech to Text

1. The goal is to generate a text from an audio signal.



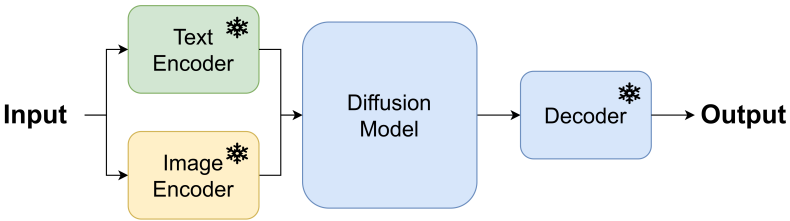
2. A great example of a Generative AI Systems for transforming speech to text is Whisper (Radford et al. n.d.).
3. Whisper uses an encoder-decoder transformer with a specific form of the encoder.
4. Whisper model is an automatic speech recognition system with
   - **a tiny version:** 39M weights
   - **a large version:** 1.55B weights

Example: Speech to Text

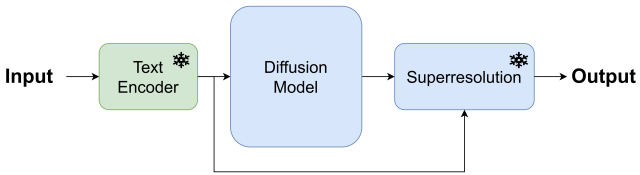


Example: Large vision models

1. Large Vision Models (LVMs) are perfect examples of Generative AI Systems:
   - **Image to text**
   - **Text to image**
2. Latent diffusion models are widely used for generating images for a given prompt (Rombach et al. 2022).



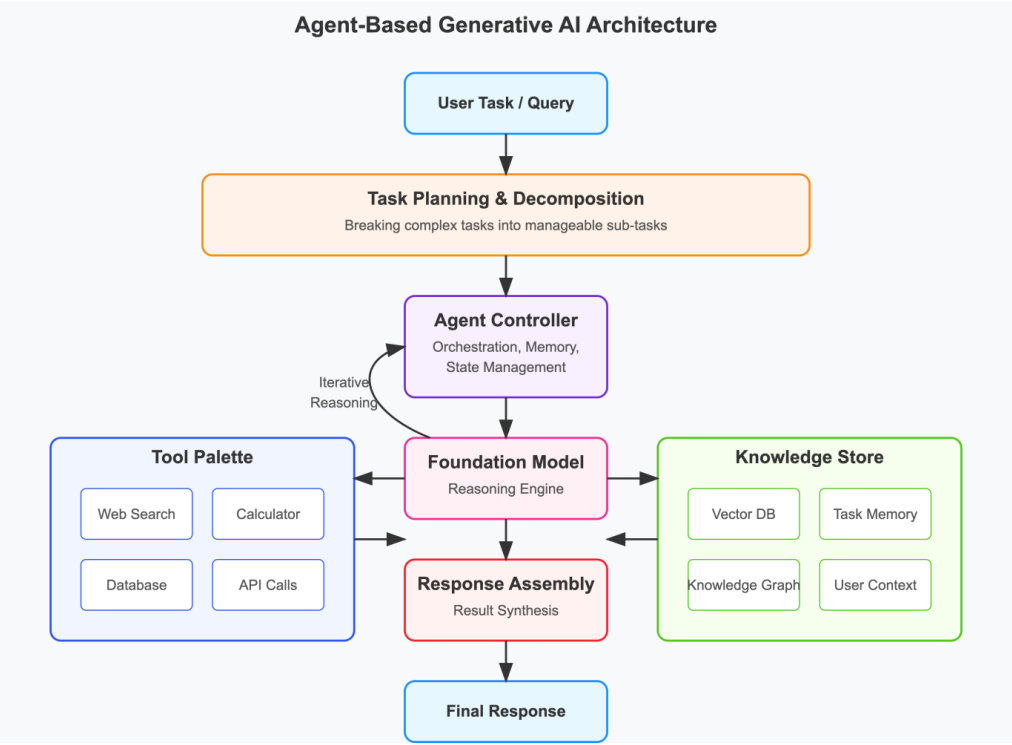
3. ImaGen uses a T5-based text encoder and a diffusion model together with superresolution blocks (Saharia et al. 2022).



Example: Agent-Based Systems

1. The idea of using LLMs as a backbone for Operating Systems and agents as applications has attracted a lot of attention.
2. Another idea that is pretty hyped these days is **Agentic AI**
   the development of GenAISys-based agents operating in an autonomous manner
   - sophisticated planning,
   - select appropriate tools for each component,
   - execute operations via well-defined APIs,
   - use interim results to inform subsequent reasoning steps, and
   - finally synthesize findings into coherent outputs.
3. This approach enables generative AI to tackle problems requiring
   - prolonged reasoning,
   - external knowledge access, and
   - specialized computational capabilities beyond what's possible with standard prompting.

Example: Agent-Based Systems



Summary

1. The idea of using LLMs as a backbone for Operating Systems and agents as applications has attracted a lot of attention.
2. Another idea that is pretty hyped these days is **Agentic AI**
   the development of GenAISys-based agents operating in an autonomous manner.
   - Microsoft AutoGen
   - OpenAI ChatGPT

References

Reading

1. Chapter 11 of Deep Generative Modeling (Tomczak 2024).

References i

- Lewis, Patrick et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*.
- Radford, Alec et al. (n.d.). "Robust Speech Recognition via Large-Scale Weak Supervision". In: *International Conference on Machine Learning*. Vol. 202, pp. 28492–28518.
- Rombach, Robin et al. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685.
- Saharia, Chitwan et al. (2022). "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". In: *Advances in Neural Information Processing Systems*.
- Tomczak, Jakub M. (2024). *Deep Generative Modeling*. Springer.
- Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Questions?