Cyber2A Workshop

Foundation Models: The Cornerstones of Modern AI

Minu Mathew
Research Software Engineer
National Center for Supercomputing Applications

Sandeep Puthanveetil Satheesan
Senior Research Software Engineer
National Center for Supercomputing Applications

import MinuMathew as mm

mm.work()

Research Software Engineer

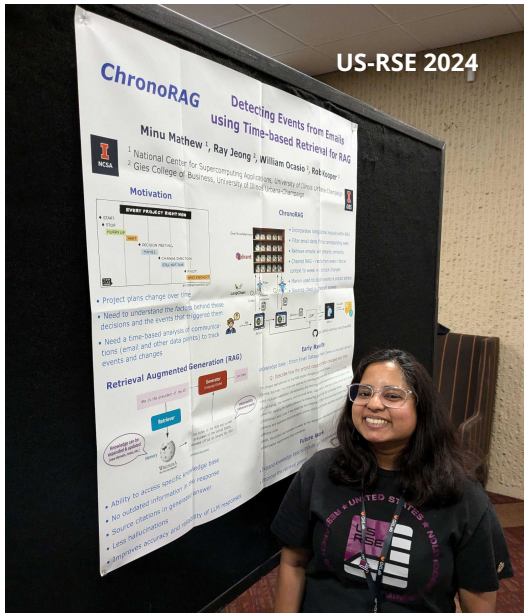
Adjunct Lecturer

Machine Learning Intern

Research Student

Software Engineer

mm.meet()



mm.fun()

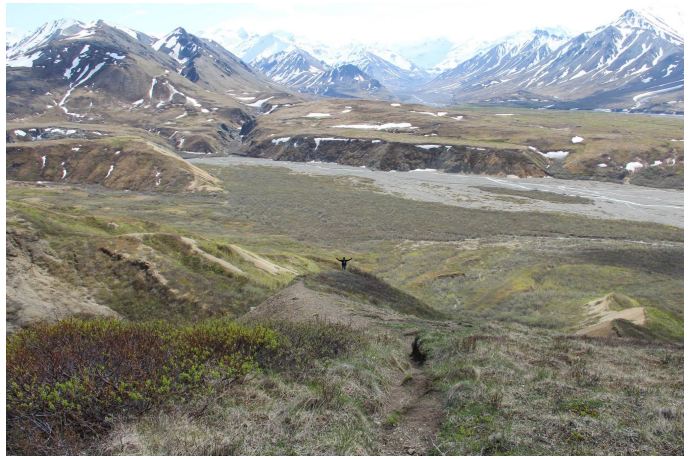
import SandeepPuthanveetilSatheesan as sps

sps.work()

- Senior Research Software Engineer — NCSA
- Senior Research Programmer — NCSA
- Research Programmer — NCSA
- Visiting Research Programmer — NCSA
- Software Developer (Student Assistant)
- Software Engineer — Infosys
- Software Engineer Trainee — AMRITA VISHWA VIDYAPEETHAM

sps.meet()

sps.fun()



Agenda

- Foundation models overview
- Types of foundation models
- Common architectures
- Segment Anything Model 2 (SAM 2)
- Retrieval Augmented Generation

Slides :

https://github.com/cyber2a/cyber2a-course/tree/main/slides

Course book :

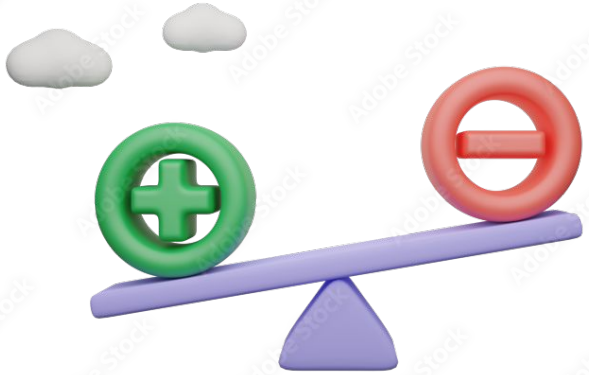
https://cyber2a.github.io/cyber2a-course/sections/foundation-models.html

Disclaimers:

- Fast evolving field
- Provided is a brief overview

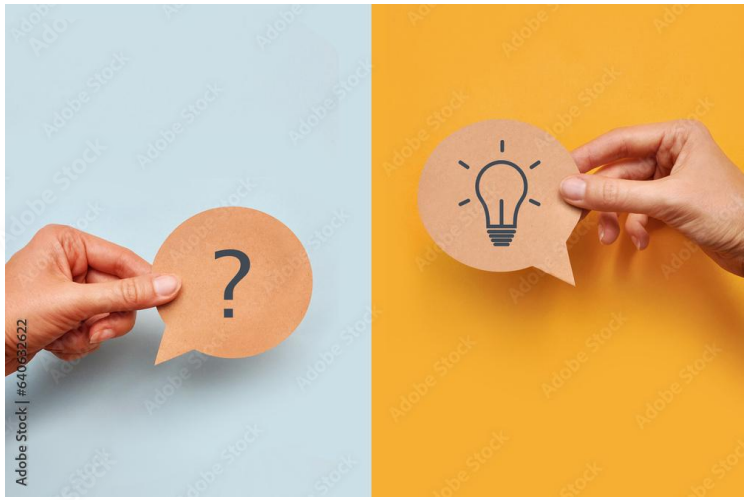
Just a few years ago, in the world of language…

Language translation Google translate



Sentiment analysis Sentence classification

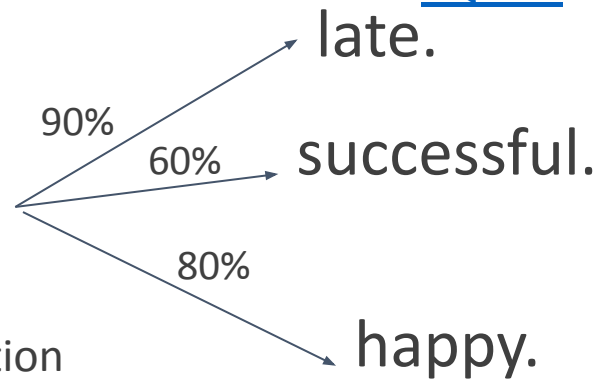
Dataset : IMDB, Amazon reviews



Question answering

Dataset : SQUAD dataset

- Different models for different tasks

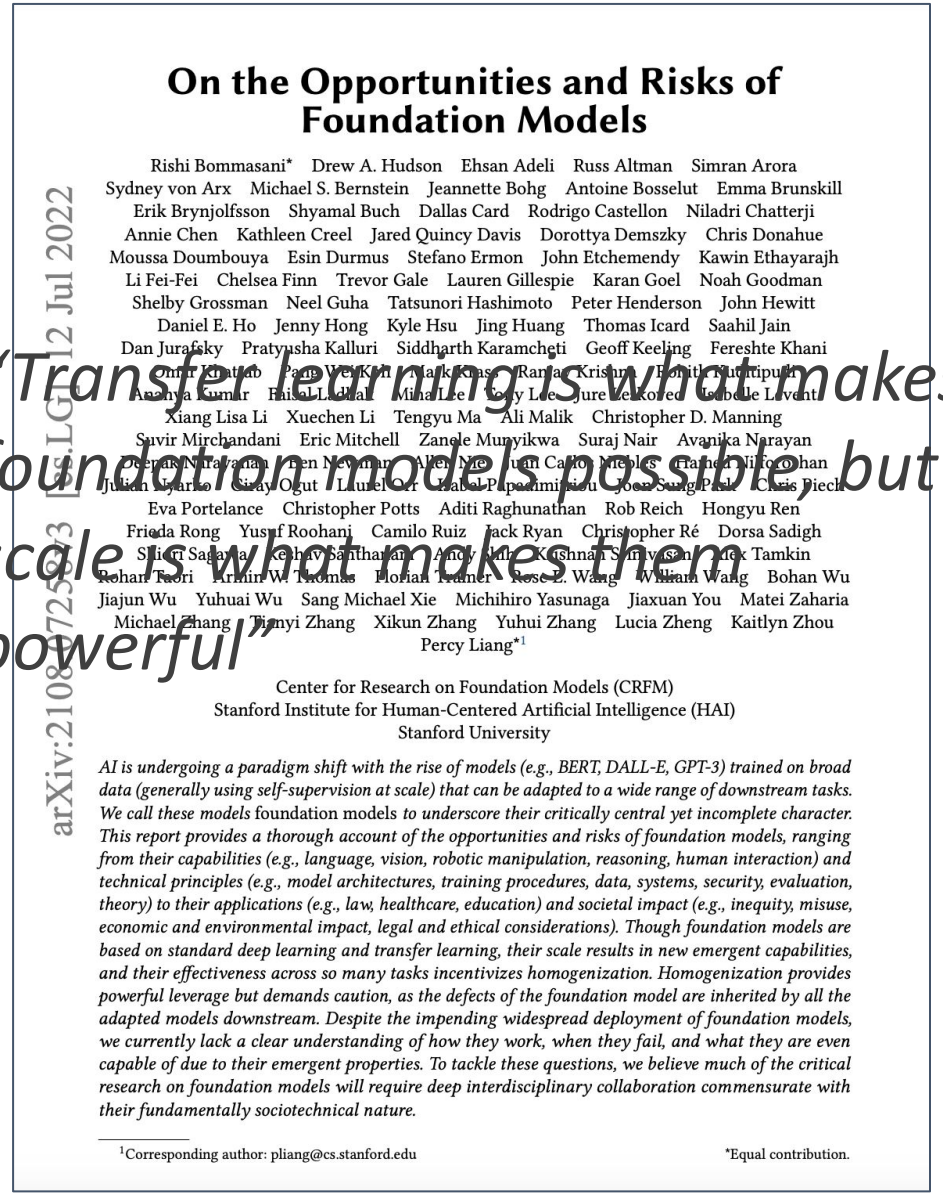


Sentence completion

Source : MIT class on foundation models

Foundation Models

- Coined in 2021 (paper)
- Paradigm shift in AI
- **Huge** neural networks
- **Vast** amount of training data
  - Typically **self-supervised** learning
  - “Intelligence” comes with more data
- **Massive** compute power for training
- **Adapted** to various downstream tasks



On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks. We call these models foundation models *to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotic manipulation, reasoning, human interaction) and technical principles (e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.*

[1]Corresponding author: pliang@cs.stanford.edu *Equal contribution.

“Transfer learning is what makes foundation models possible, but scale is what makes them powerful”

Foundation Models

- Serves as a “foundation” from which many task-specific models can be built by adaptation

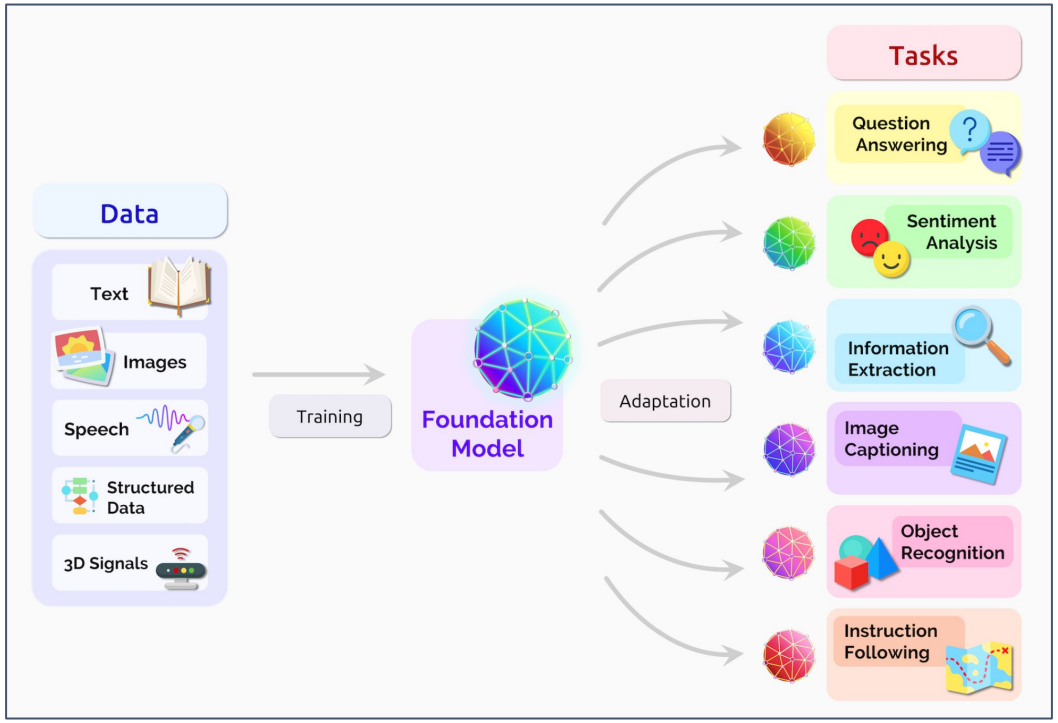


Image credits: Foundation Models paper

- With the right model architecture, intelligence comes with large-scale data
- Examples:
  - CLIP - 63 million parameters
  - BERT - 345 million parameters
  - GPT-3 - 175 billion parameters
    - Wikipedia = 3% of the total training data
  - GPT-4 - 1.8 trillion parameters

Types of Foundation Models

Criteria: Modality	Criteria: Architecture
Language Models	Transformer Models
Vision Models	Generative Models
Multimodal Models	Diffusion Models

Types of Foundation Models (Modality)

- **Language models**
  - Trained on textual data
  - Good at translation, conversational AI, sentiment analysis, content summarization, etc.
  - Eg: LLMs - GPT-3, GPT-4, Llama 3.2



- **Vision models**
  - Trained to include computer vision tasks
  - Good at Object detection, Segmentation, Facial recognition, etc.
  - Eg: GPT-4-turbo, SAM, Swin-Transformer



Types of Foundation Models (Modality)

- **Multimodal models**
  - Trained on datasets that include multiple modalities
    - Learn relationships and correlations between different forms of information
    - Eg: GPT-4o can reason across text, vision and audio
  - Contrastive learning
    - Understand how different modalities relate.
    - Aim to maximise the similarity between paired data (eg image and its caption) while minimizing similarity between unrelated pairs.
  - Eg: DALL-E, GPT-4o, CLIP, Sora, Gemini

Types of Foundation Models (Architecture)

Transformer

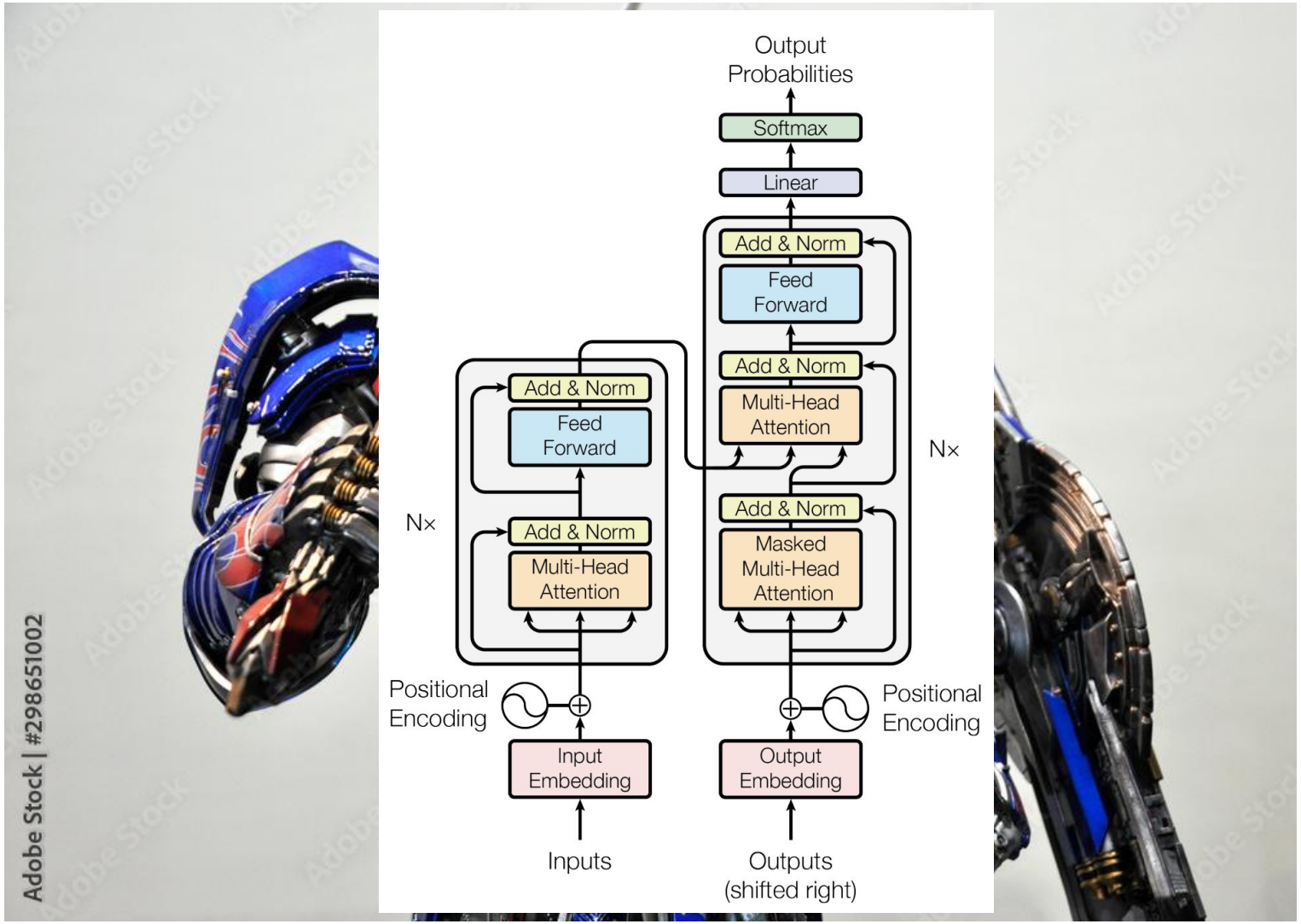


Image source: https://arxiv.org/html/1706.03762v7

Types of Foundation Models (Architecture)

- **Transformer models**
  - Introduced in 2017 "Attention is all you need"
  - Key features -
    - attention mechanism, positional encodings
  - Eg: GPT-3, CLIP
- **Generative Adversarial models**
  - Generative Adversarial Networks introduced in 2014
  - Involves a generator-discriminator network pair which competes with one another.
  - Eg: StyleGAN, BigGAN
- **Diffusion models**
  - Introduced in 2020 paper
  - Training involves adding random noise in steps and learning to remove the noise
  - Eg: Stable-diffusion, DALL-E, Sora

Types of Foundation Models (Architecture)

Transformers

- Proposed in 2017 paper: “Attention Is All You Need”
- Blog post , video1, video2 on transformers
- Two blocks - Encoder and Decoder
  - Encoder extracts features from input
  - Decoder uses the features to produce the output

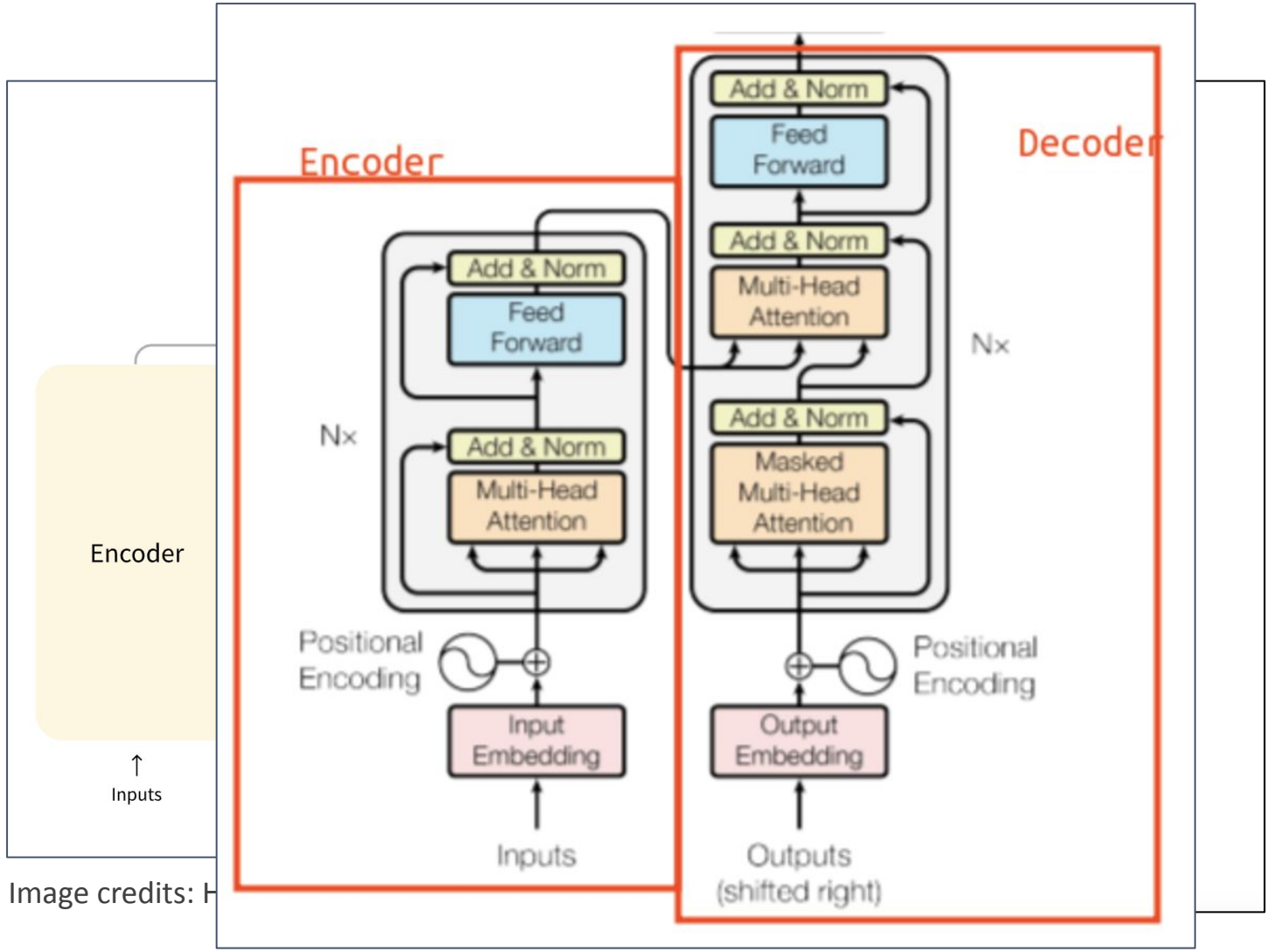


Image credits: H

Image source: https://arxiv.org/html/1706.03762v7

Types of Foundation Models (Architecture)

Transformers

- Attention mechanism (blog post)
- Positional encodings
- More than 50 major transformer models

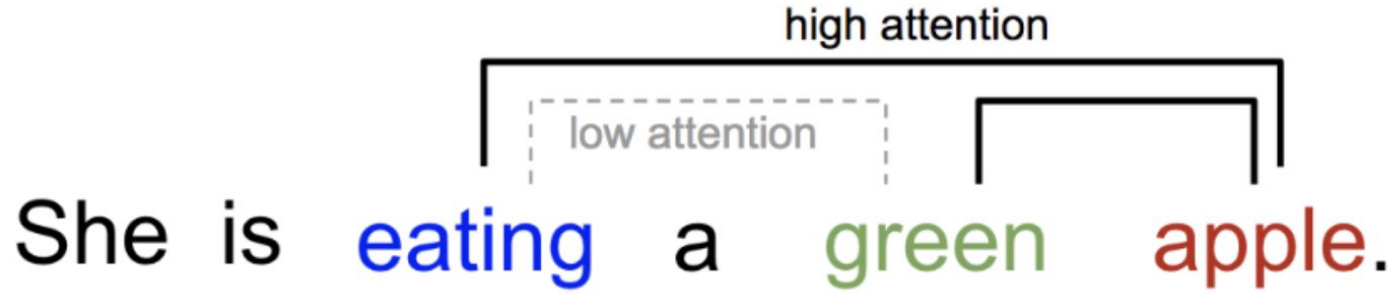


Fig : One word attends to other words in the same sentence differently

Image credits: LilanWeng's blog

Types of Foundation Models (Architecture)

- **Generative Adversarial Networks (GAN)**
  - Good at generating synthetic data: Photo editing, face aging, and different human poses.

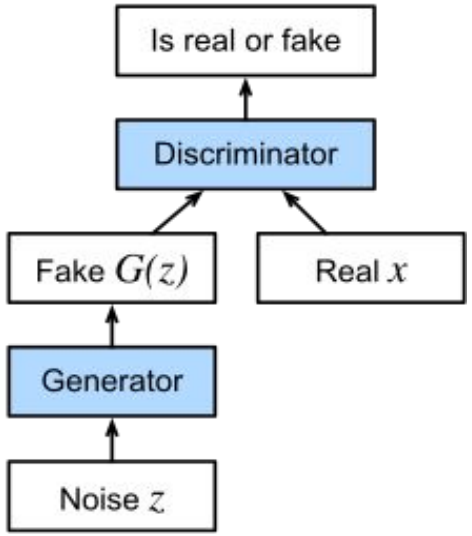


Image credits: Wikipedia

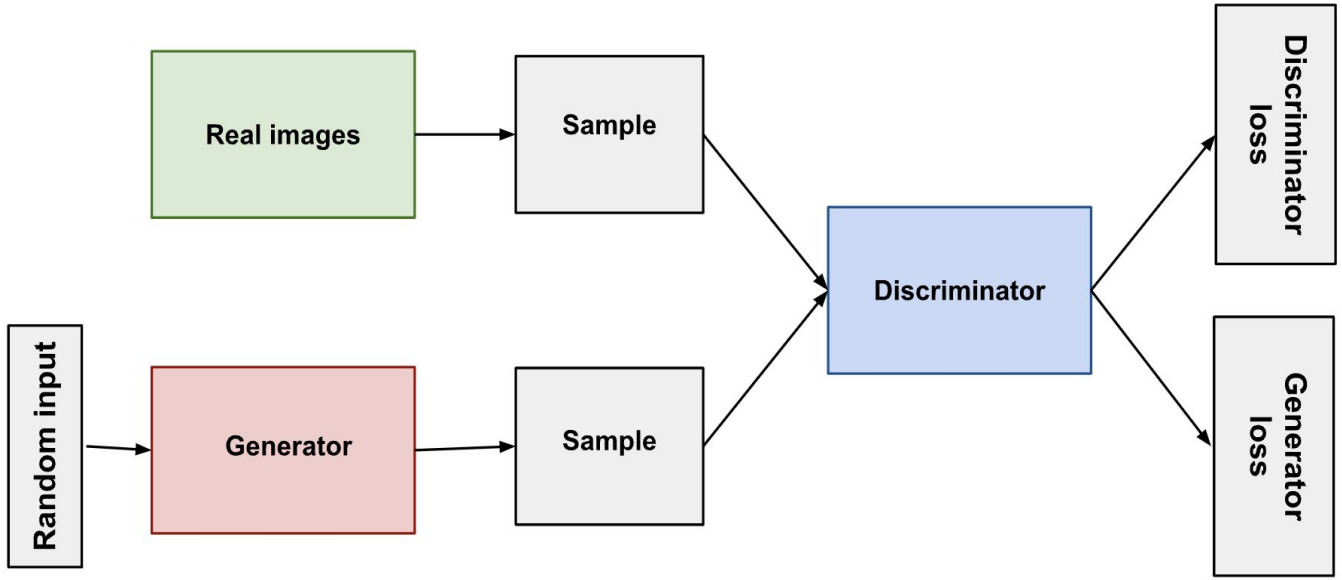


Image credits: Google developers blog

  - Involves a generator(G)-discriminator(D) pair.
  - Training procedure for G is to maximise the probability of D making a mistake.

Types of Foundation Models (Architecture)

- **Generative Adversarial Networks (GAN)**
  - Unstable network : model convergence is difficult.
  - Less diversity in generated outputs.

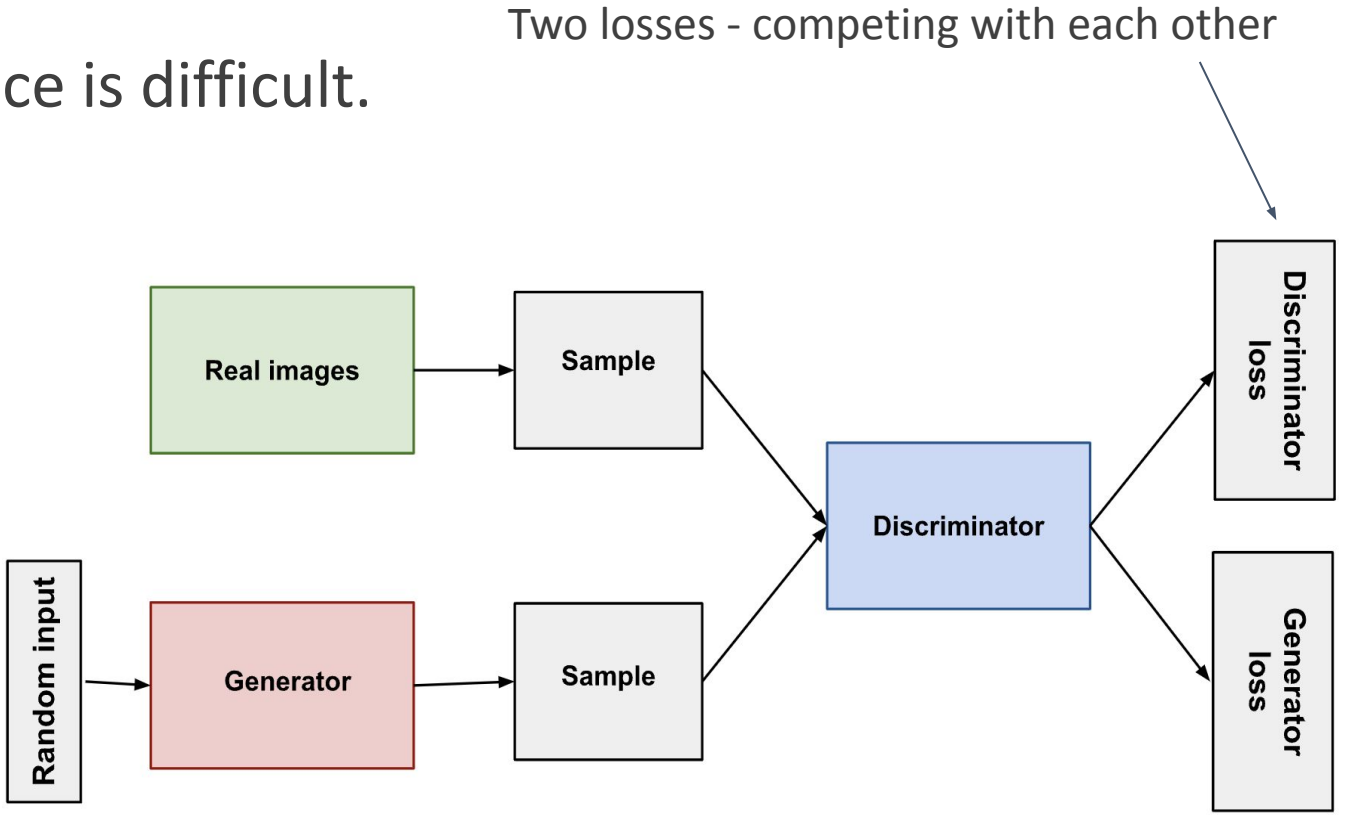


Image credits: Google developers blog

Types of Foundation Models (Architecture)

- **Diffusion models**
  - Learns by adding and removing noise
  - Eg: DALL-E, Stable diffusion, Sora
  - More stable than GANs
  - More diverse outputs
  - Blog post on diffusion models

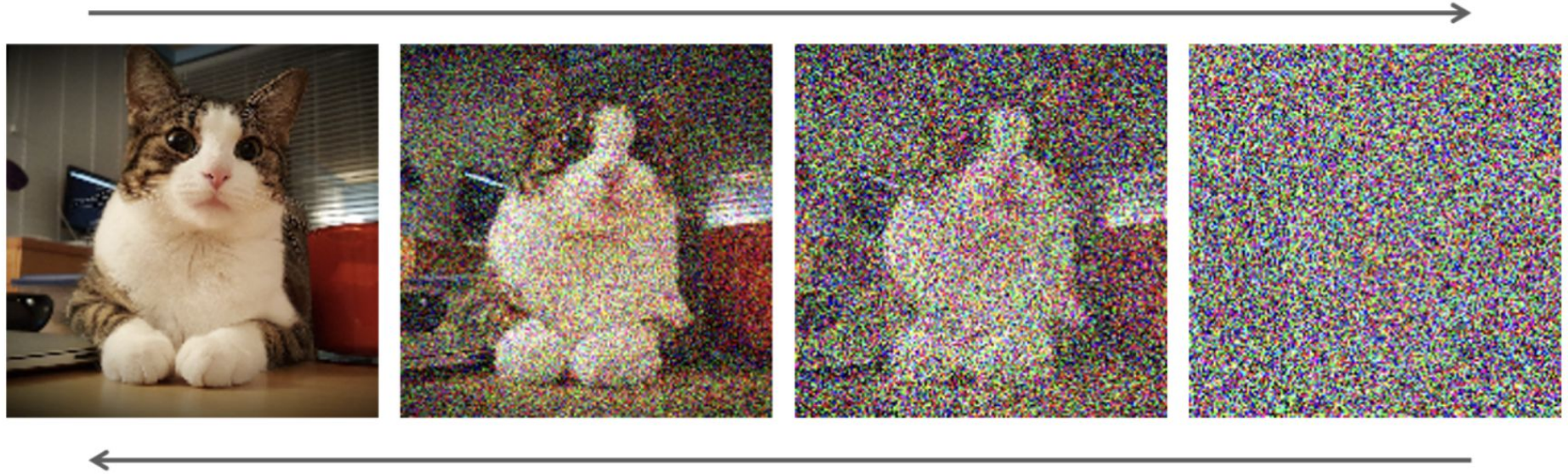


Image Source: nvidia blog

Segment Anything Model 2 (SAM 2)

Overview

SAM 2 Overview

- Developed by Meta Segment Anything project
  - Segment Anything Model - introduced a foundation model for promptable segmentation task (2023)
  - Defined a task (promptable segmentation), model, and dataset
  - SA-1B dataset: ~11 million images and ~1 billion image masks
- SAM 2 (Ravi et al., 2024)
  - Unified model for promptable visual (image and video) segmentation
  - SA-V dataset: ~50K videos and ~640K masklets (*spatio-temporal masks*)
  - Prompts: **Points**, **Bounding Box**, or **Mask**

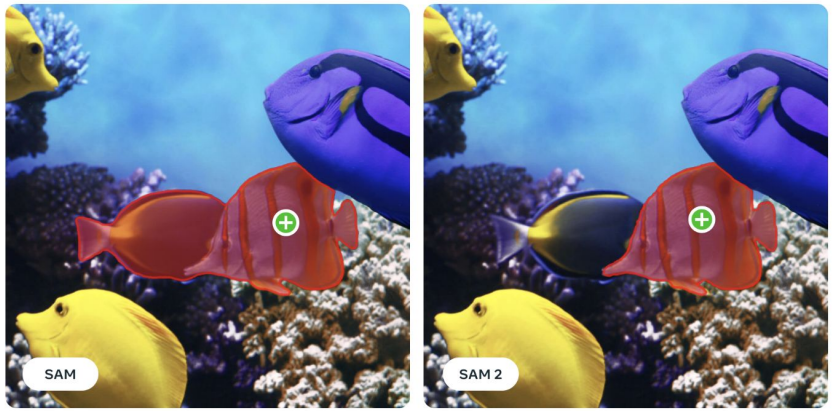


Image and Video source: Meta's blog on SAM 2

SAM-2 Overview

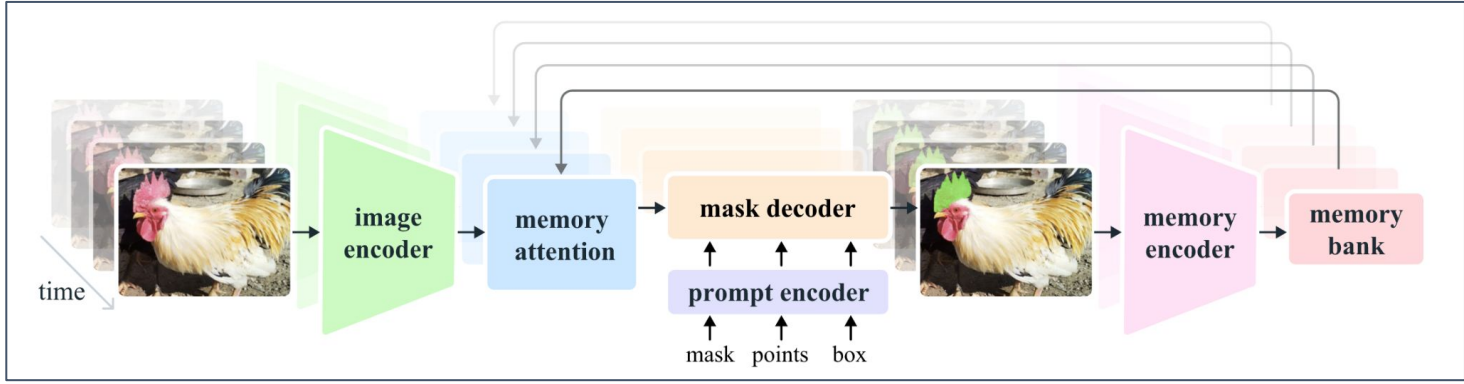


Image source: [Ravi et al., 2024]

- For segmenting both videos and image (i.e., single-frame video)

- **Image encoder**
  - Provides feature embeddings for each frame

- **Memory attention**
  - *conditions* current frame features on past frames features, predictions, and any new prompts

- **Prompt encoder**
  - Supports sparse (points and boxes) and dense (mask) prompts
  - Represented by different forms of encodings/embeddings.

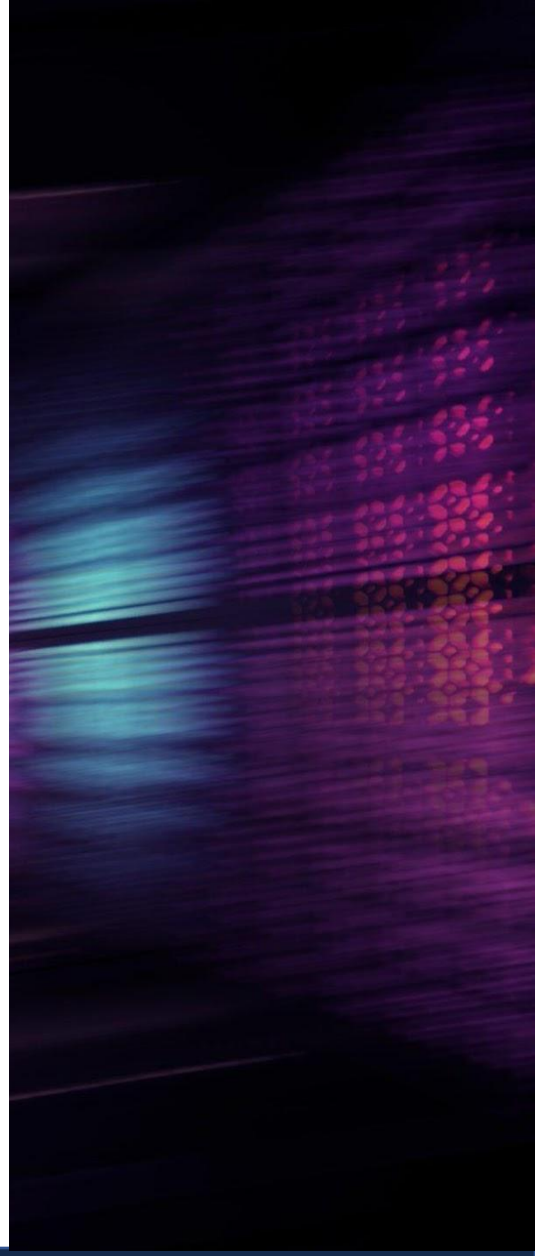
- **Mask decoder**
  - current frame image embedding + encoded prompts = segmentation mask

Retrieval Augmented Generation

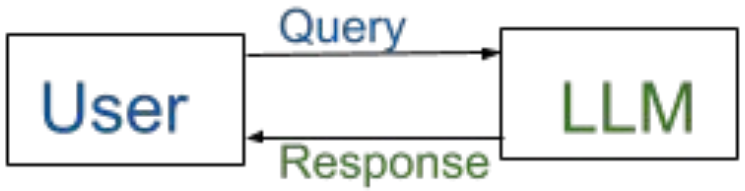
Overview

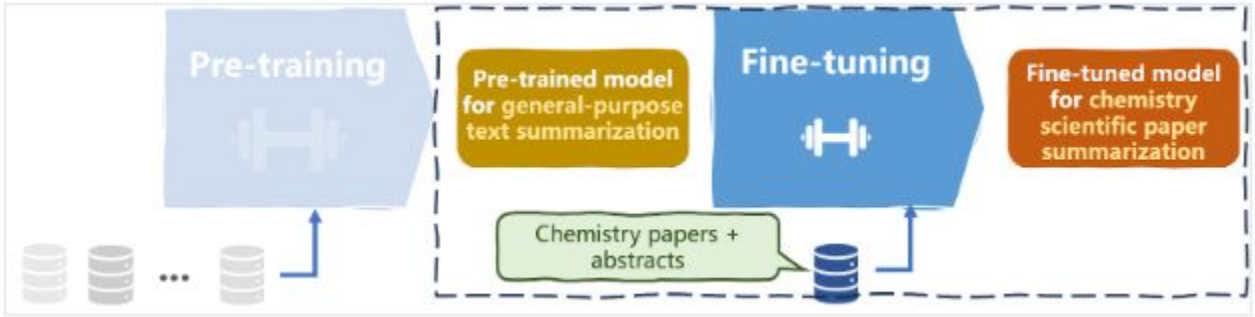
Motivation

- LLM - Large pre-trained Language Models :
  - No ability to access a specific knowledge
    - LLM has not seen “your” data.
    - If specific domain knowledge is required - fine tune.
  - No sources
  - Hallucinations
  - Out-of-date information



Motivation





Fine-tuning LLM
Source: [datacamp blog]

Explain to me what an llm is

An LLM, or Master of Laws, is a postgraduate degree in law that is typically pursued by individuals who have already completed a bachelor's degree in law (LLB or JD) or have significant legal experience. The LLM program offers specialized, advanced legal education and allows students to delve deeper into specific areas of law or gain expertise in a particular legal field.

When did France gift Lithuania Vilnius TV tower?

France gifted Lithuania the Vilnius TV tower in 1980.

What prompted the French government to provide the gift?

The French government provided the gift as a gesture of friendship and solidarity with the Lithuanian people, who had recently declared independence from the Soviet Union.

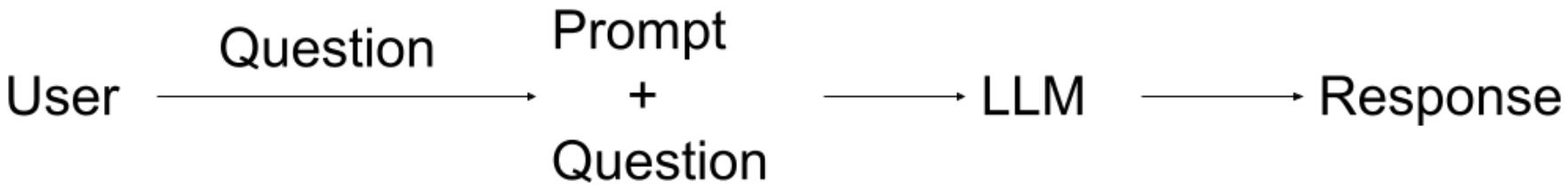
LLM Hallucinations

Introduction

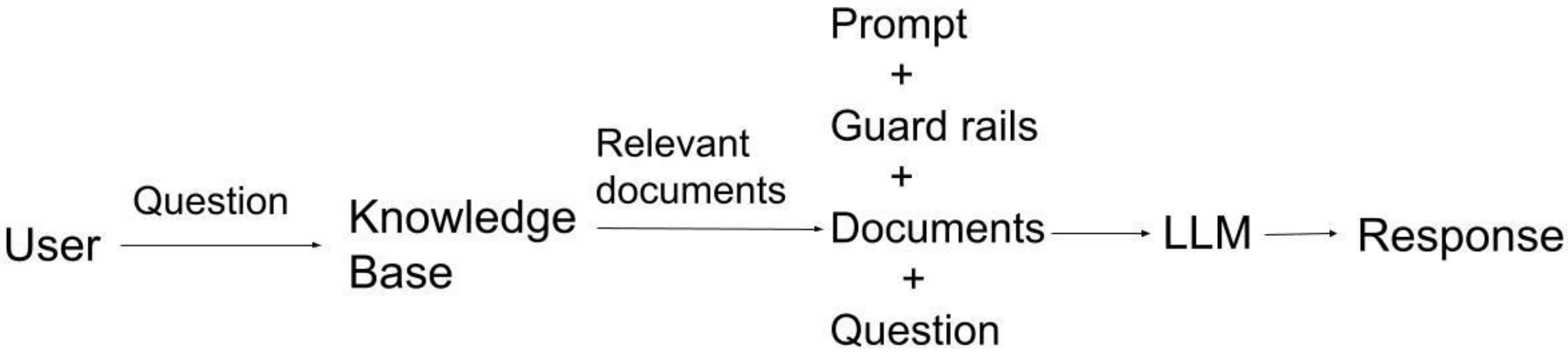
- RAG (Retrieval-Augmented Generation)
  - Introduced in 2020 (paper)
  - Generation - by LLMs
  - Retrieval-Augmented
    - Retrieve required information from provided knowledge base
    - Provide this information to LLMs as context to answer user query

Introduction

- Before :



- After :



Without RAG	With RAG
• No ability to access a specific knowledge/domain	• Point to a knowledge base
• No sources	• Sources cited in LLM response
• Hallucinations	• LLM response is grounded by relevant information from knowledge base
• Out-of-date information	• Update the knowledge base with new information

Acknowledgement

- This material is based upon work supported by the National Science Foundation under Grant Nos. #2230034 and #2230035.
- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

- We want to thank Rob Kooper, Lead Research Software Engineer, National Center for Supercomputing Applications for setting up the Jupyter Server used for the MLFlow and Foundation Models hands-on sessions.

That’s all folks!

Minu Mathew
minum@illinois.edu
www.linkedin.com/in/minumpmathews

Sandeep Puthanveetil Satheesan
sandeeps@illinois.edu
https://sandeep-ps.github.io/

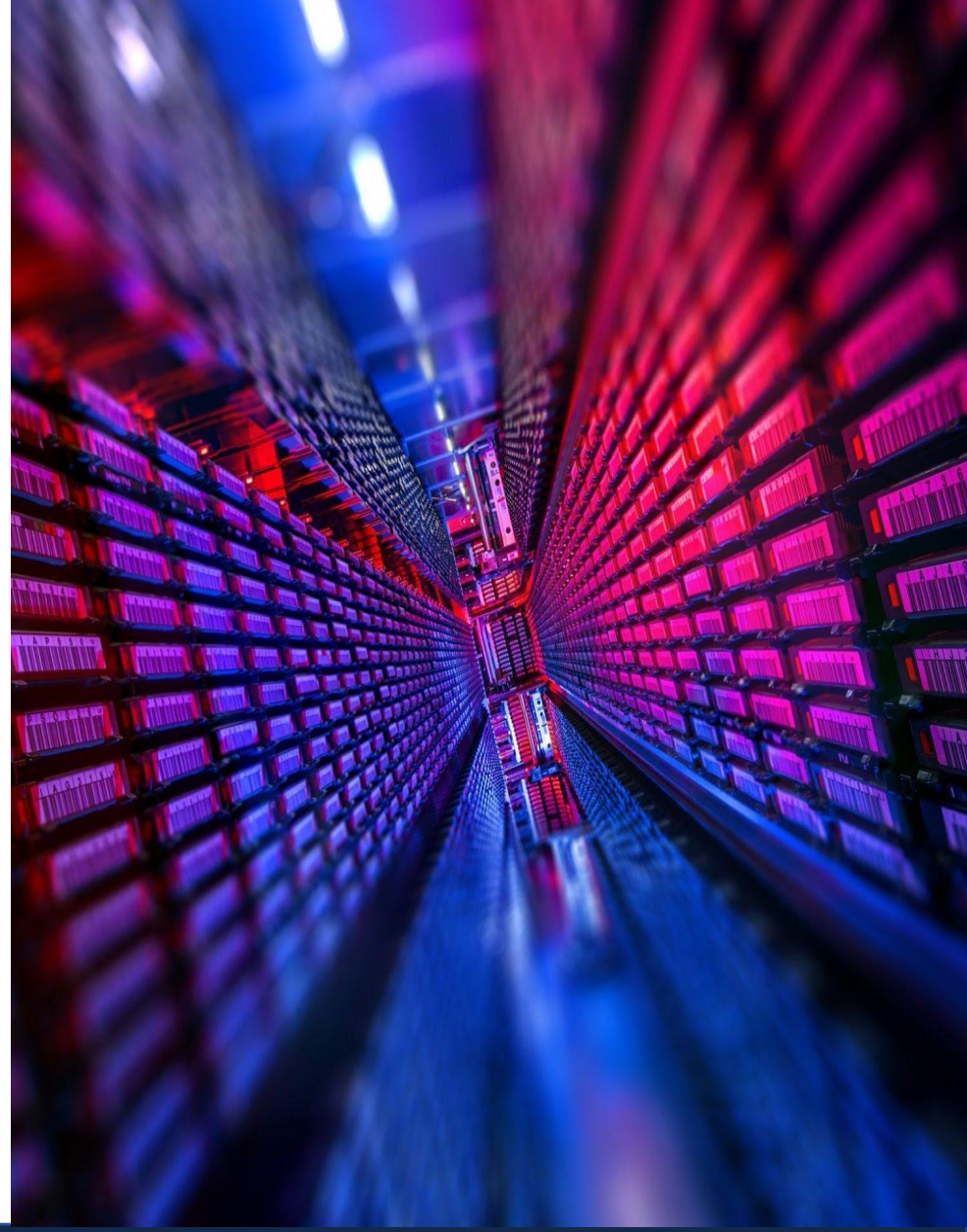
Foundation Models Hands-on

cyber2a.software.ncsa.illinois.edu

github.com/ncsa/cyber2a-workshop

Hands-on

- Github
  https://github.com/ncsa/cyber2a-workshop
- Jupyter hub
  https://cyber2a.software.ncsa.illinois.edu/
- Details also on the course book

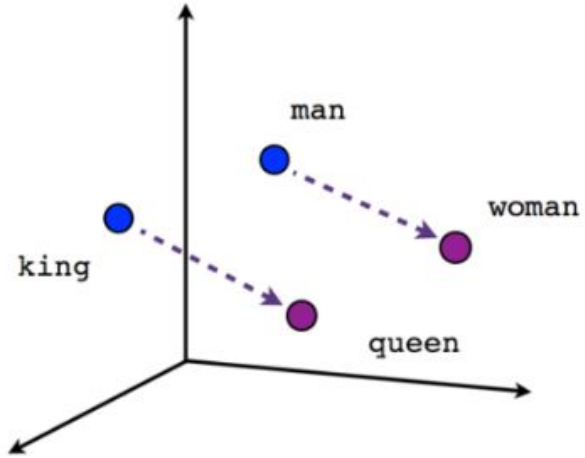


Segment Anything Model

- Promptable Visual Segmentation in Images and Videos
- Tutorial adapted from demo Jupyter Notebook shared with the SAM 2 source code.
-

RAG - Retrieval-Augmented Generation

Knowledge DB

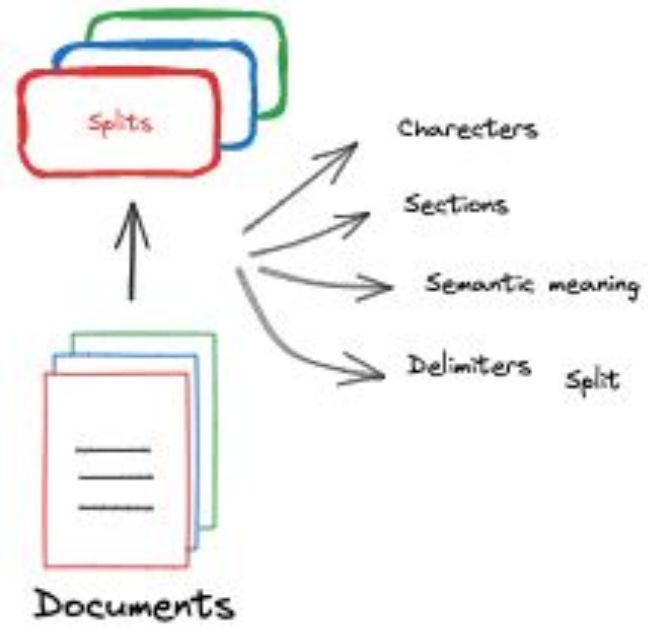


- Vector database (Beginners blog 1, Pinecone blog 2)
- Efficiently store, index and search high-dimensional data
- Store data using vector embeddings
  - Embeddings - vector representation of text.
  - Individual words are represented has real-valued vectors.
  - Captures semantic meaning and relationships of the text in a high-dimensional space.
  - Words that have similar meaning have similar representation.
- Optimized for fast retrieval and similarity search
- Calculate the distance between user query embedding and other data points

RAG - Retrieval-Augmented Generation

Chunking

- Involves breaking large amounts of data into smaller, more manageable pieces.
  - LLMs have a limited context window and can't take in the entire dataset at once.
    - For GPT-4 128k token limit.
- Try to split meaningfully
  - by pages, then by sections, paragraphs, sentences, characters..



RAG - Retrieval-Augmented Generation

Vector DB Retrieval

1. Partition data into chunks (to be processed by embedding models)
2. Use embedding model to create vector embeddings for the data (create indexes)
3. Insert vector embeddings into the db, with some reference to the original content/metadata
4. User query - use same embedding model to create user query embedding
5. Query the db for similar vector embeddings
6. Return similar document chunks.

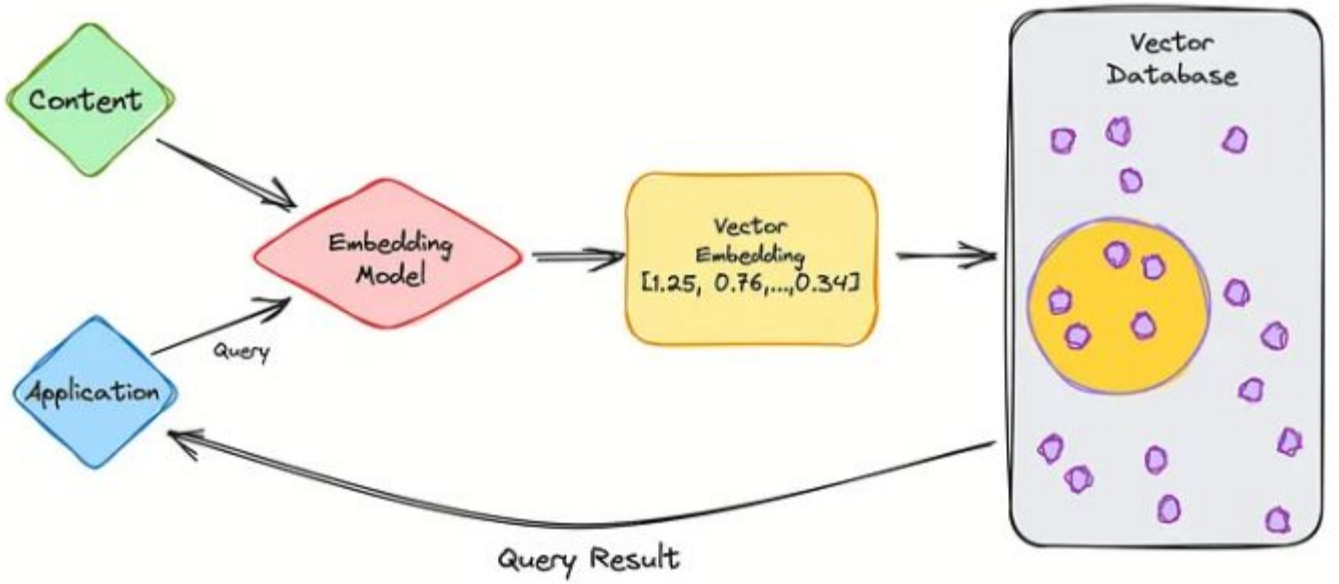


Image Credits : KDnuggets

RAG - Retrieval-Augmented Generation

LLM

- Pre-trained transformer models
- Trained to predict the next word (token), given some input text.
- Autoregressive generation - iteratively calling the model with its own generated output.
- Open-source models - Hugging Face leaderboard
- For HandsOn - GPT-4o-mini, and Llama 3

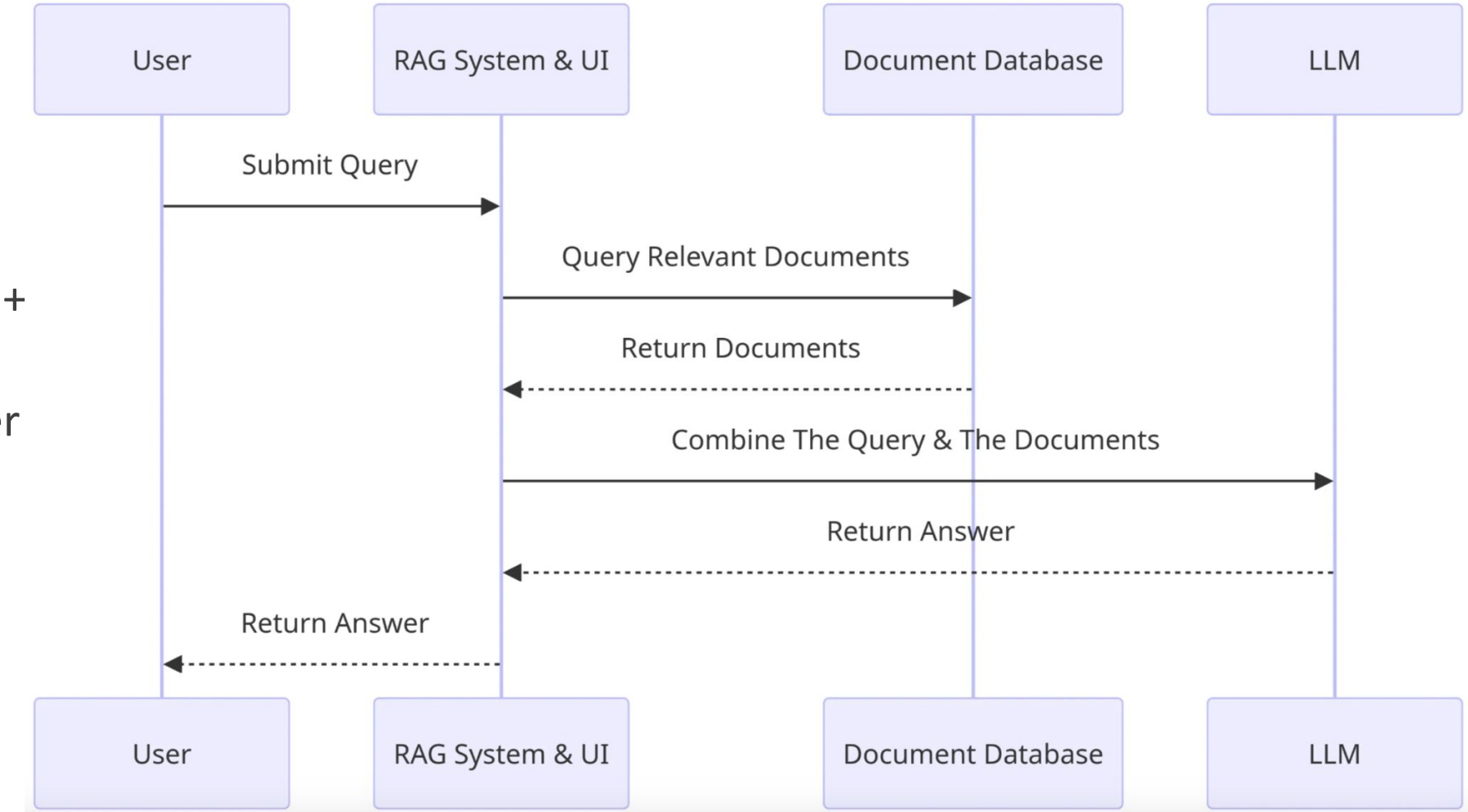
RAG - Retrieval-Augmented Generation

Prompting

- Communicate with LLMs to get desired outcomes without updating the model
- Takes multiple trial-and-errors to get desired effect.
- Include specific persona / behaviour
  - Eg: “You are a helpful research assistant”
- Include guard rails
  - Eg: “If you don’t know the final answer, just say “I don’t know”
- Include instructions
  - Eg: “Read the data file before answering any questions”
- Include response formats
  - Eg: “Respond using markdowns”
- LilianWeng blog post, medium blog post on prompt engineering

RAG - Retrieval-Augmented Generation

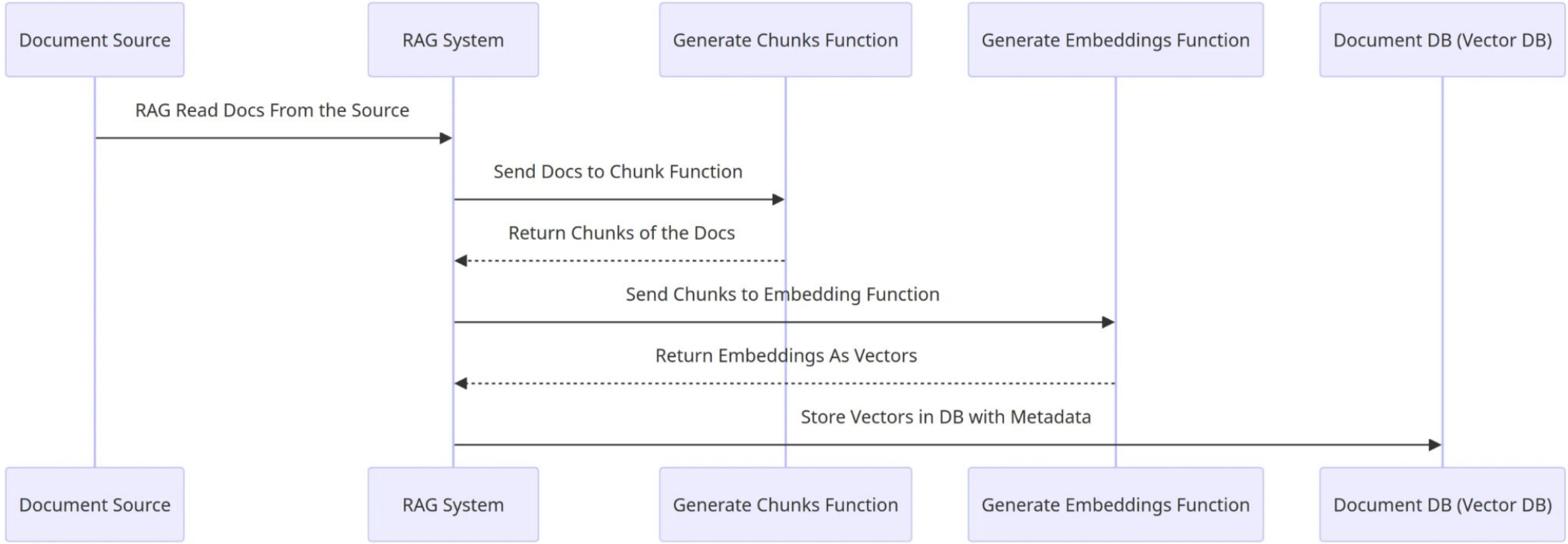
1. Query the database
2. Retrieve relevant information
3. Pass relevant documents + user query to LLM
4. Pass LLM response to user



Source Credits : [Blog.demir]

RAG - Retrieval-Augmented Generation

Inserting into DB



Source Credits : [Blog.demir]

Farewell !

Minu Mathew

minum@illinois.edu
www.linkedin.com/in/minumpmathews

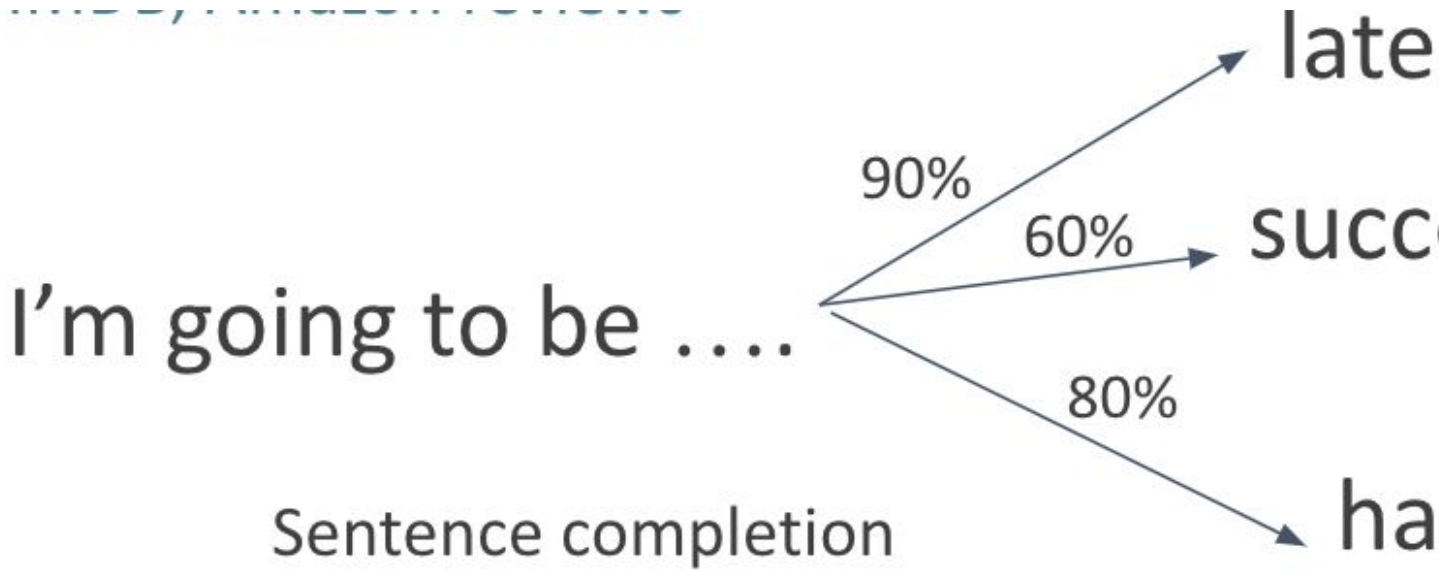
Sandeep Puthanveetil Satheesan

sandeeps@illinois.edu
https://sandeep-ps.github.io/

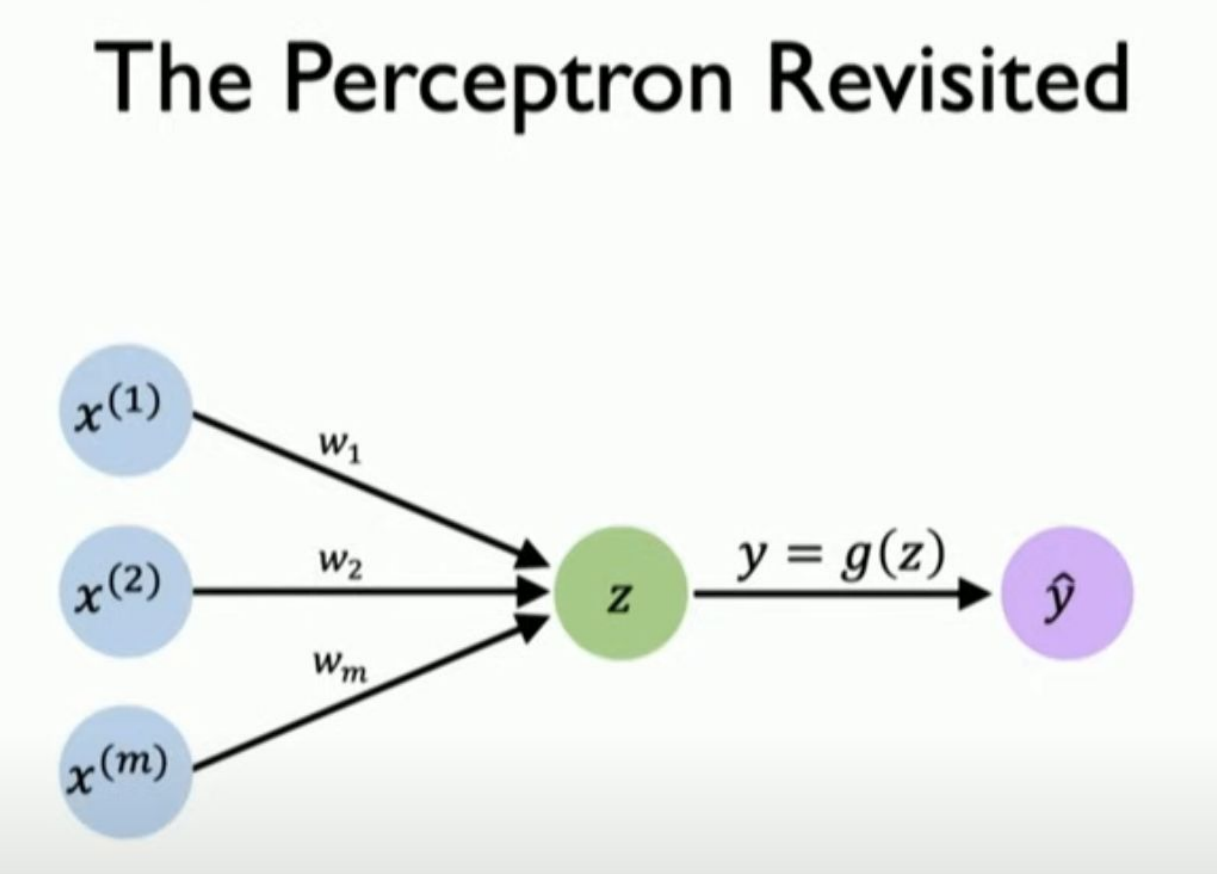
Appendix

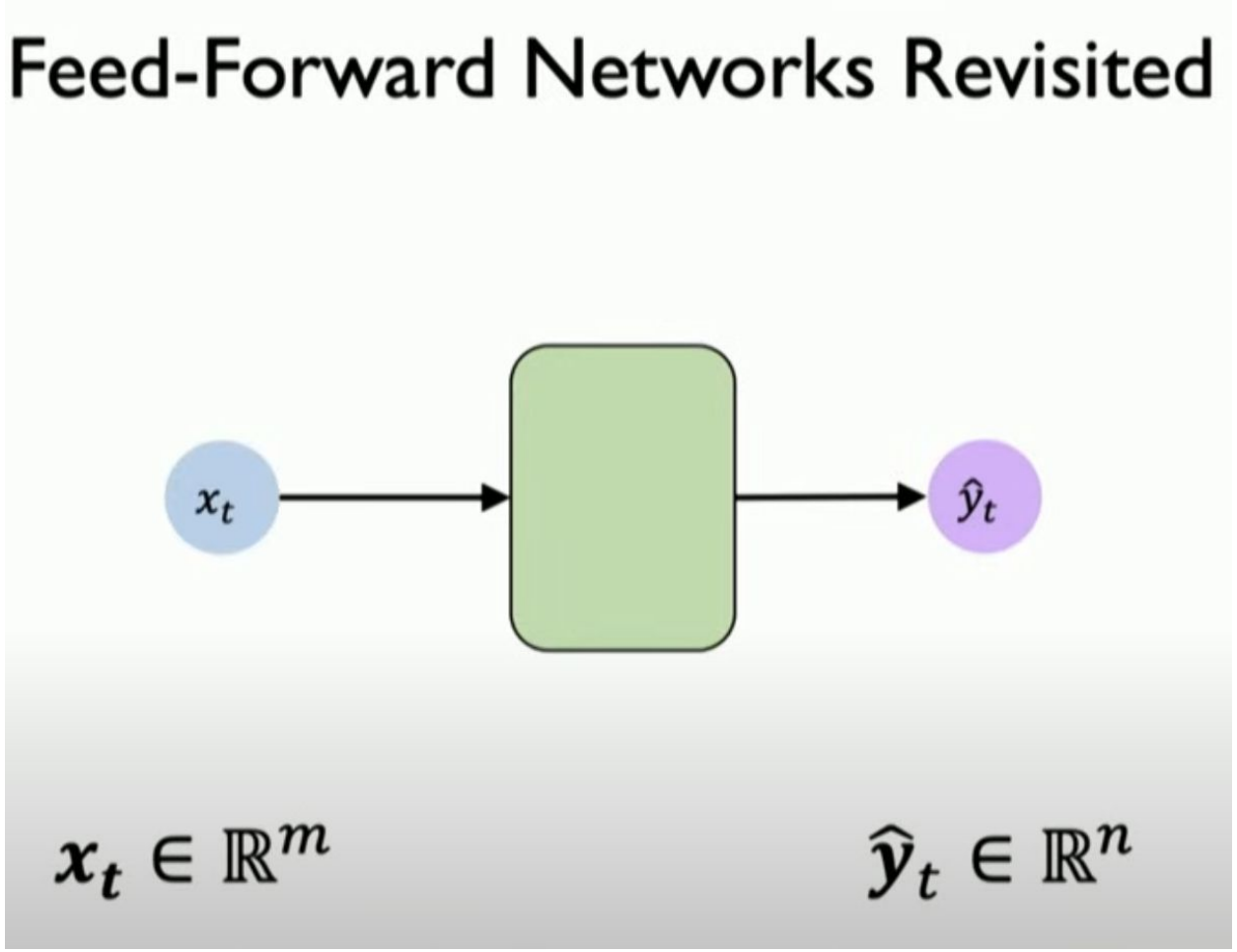
Attention Mechanism - Backdrop

Seq2Seq models



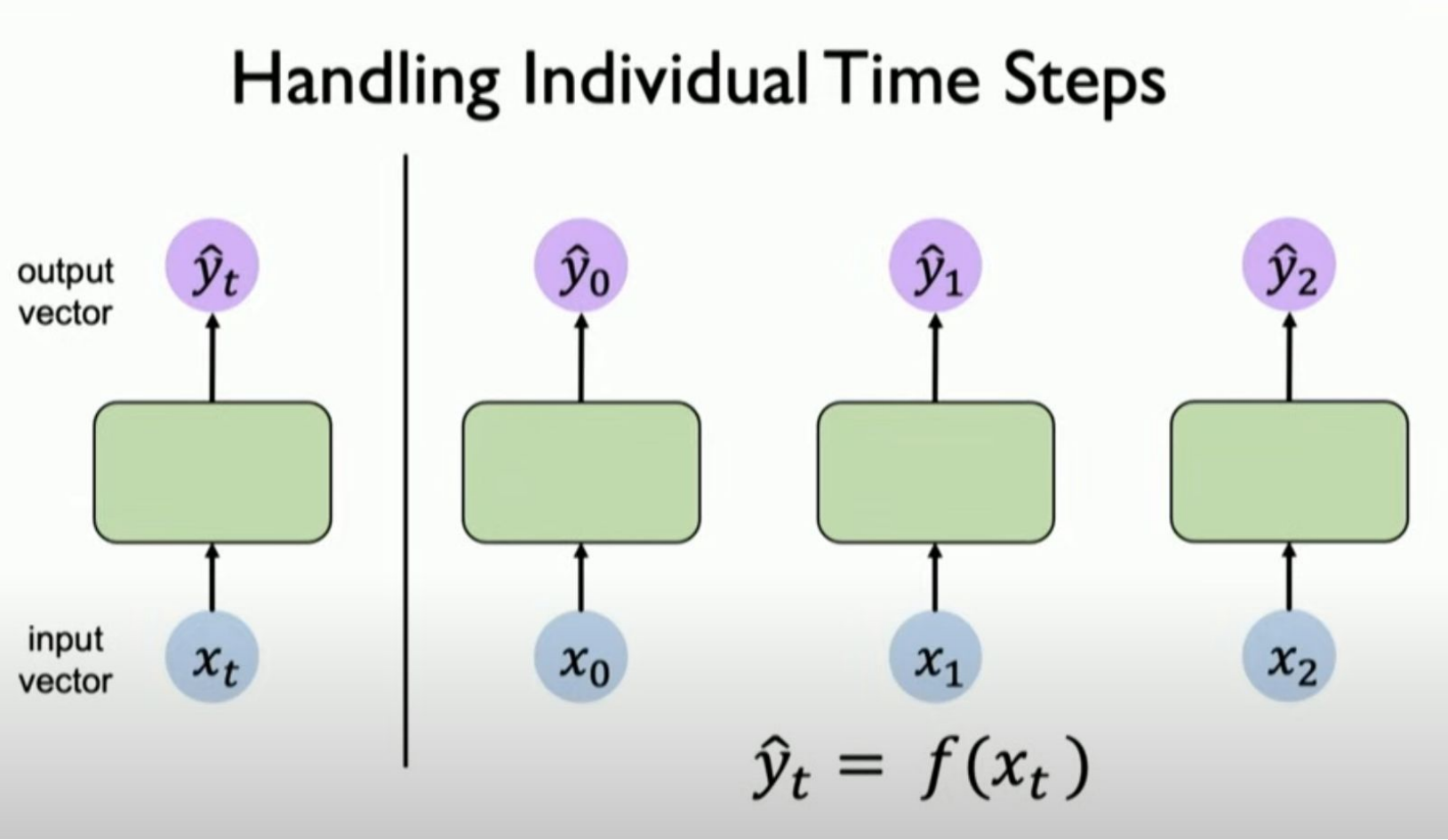
Attention Mechanism - Backdrop





Source: MIT class video

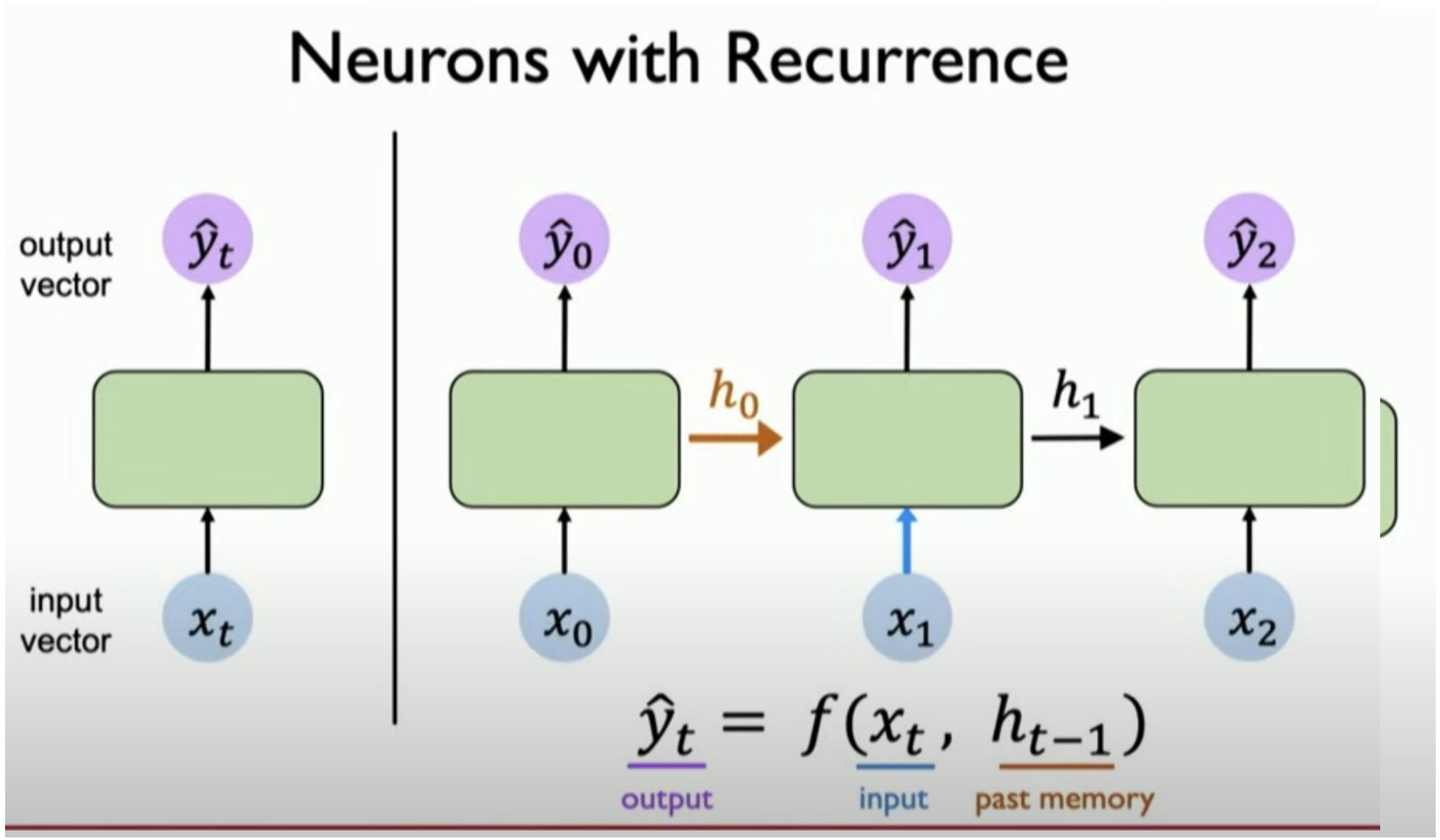
Attention Mechanism - Backdrop



Output at time t depends on input at time t-1

Source: MIT class [video]

Attention Mechanism - Backdrop



Output at time t depends on input at time t-1

Information learned at time t is passed on to t+1 using H (hidden state)

Source: MIT class video

Attention Mechanism - Backdrop

Seq2Seq models

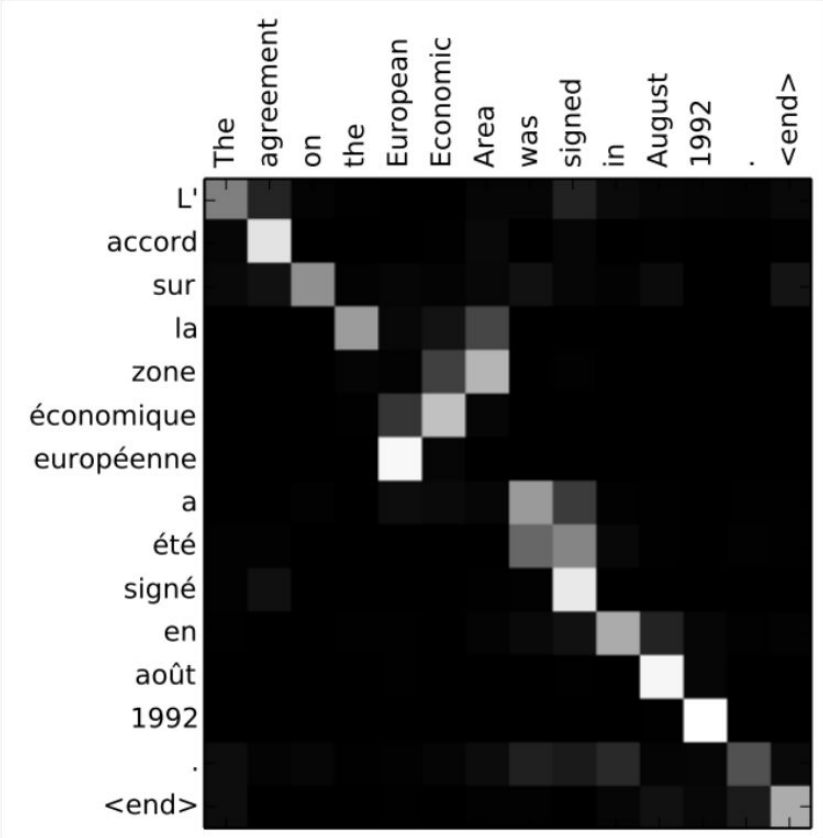
- Handle **variable-length** sequences
- Track **long-term dependencies**
- Maintain information about **order**
- **Share parameters** (hidden state information) across the sequence.
- **No parallel processing**

Link to MIT Deep Learning course slides

Foundation models - Transformers

Attention mechanism

- Introduced in Bahdanau et al., 2014 and Luong et al., 2015
- Allows the model to focus on the relevant parts of the input sequence.



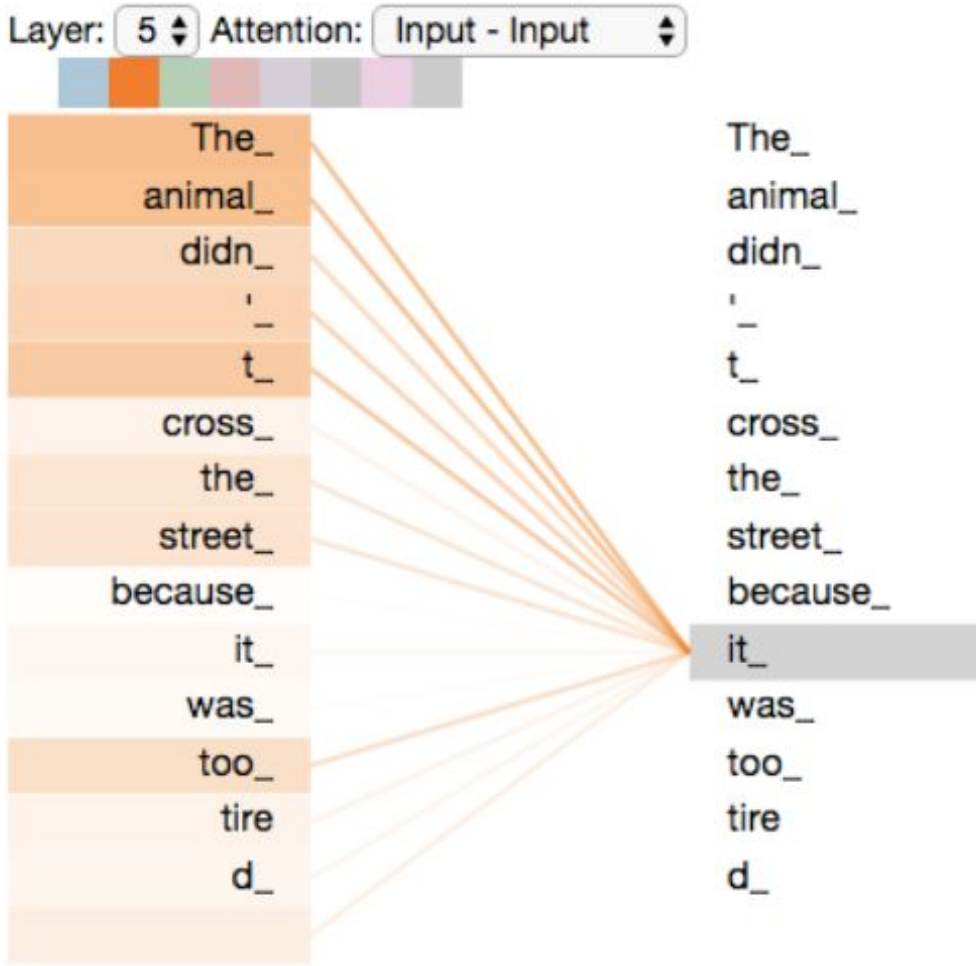


Image credits: Jay Alammar blog

Foundation models - Transformers

Positional encodings

- Order of words in the input sequence
- Add a vector to each input embedding

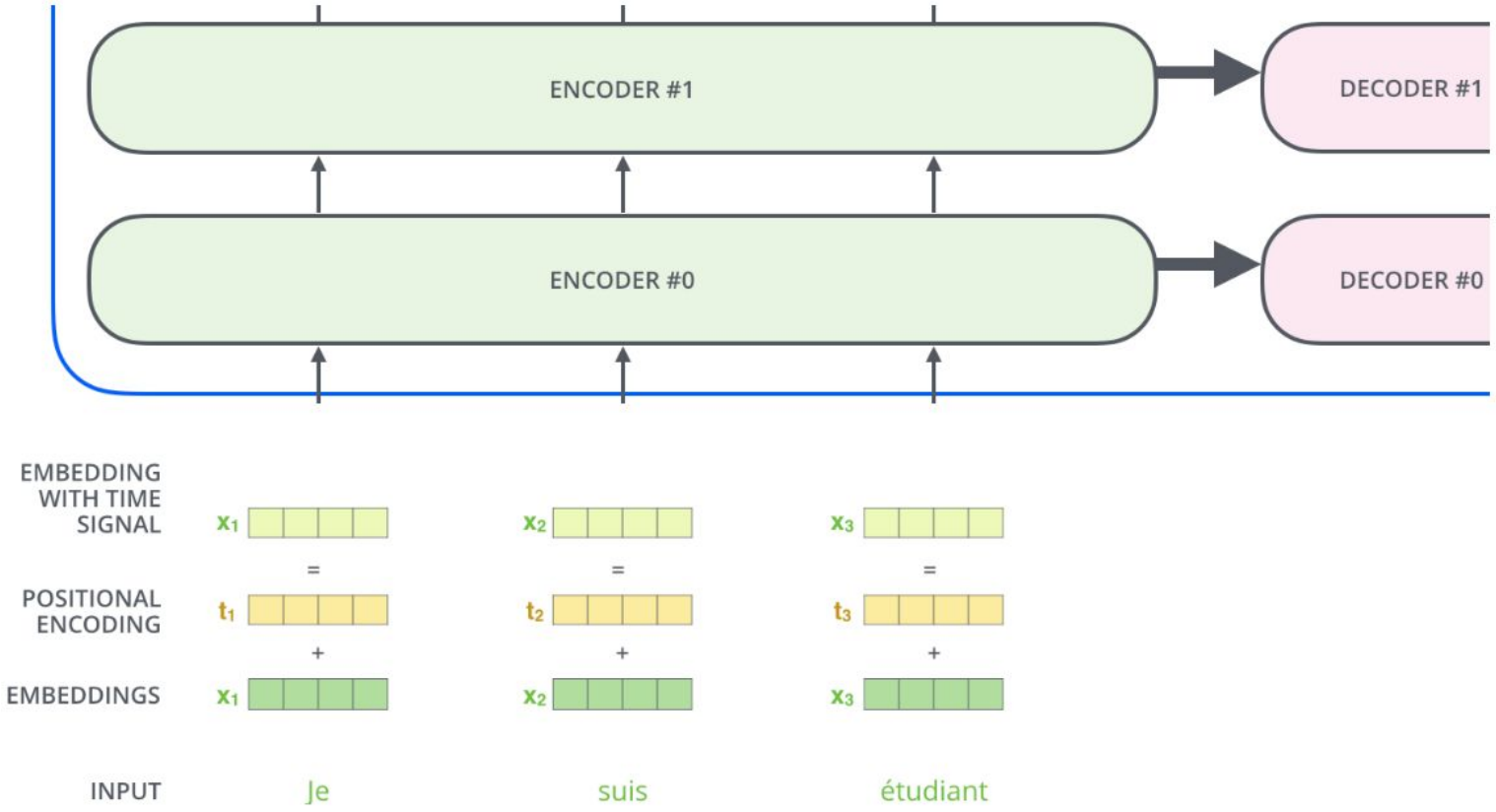


Image credits: Jay Alammar blog

Foundation models - Transformers

- Transformer architecture
- Detailed blog :

  Jay Alammar - The Illustrated Transformer
- Attention is all you need - video masterclass