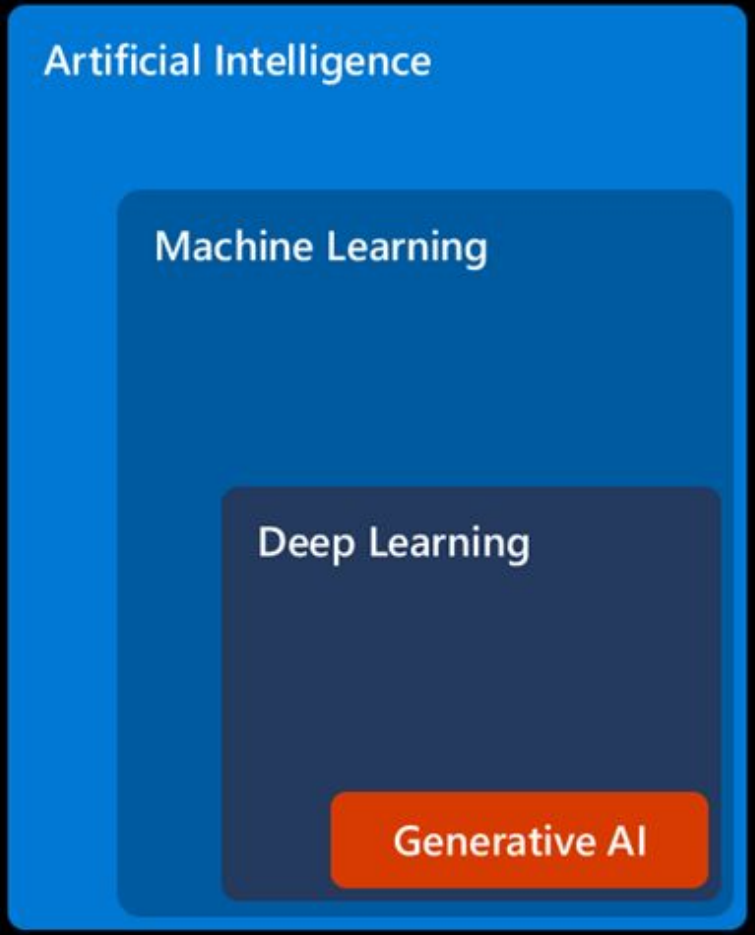
# Generative AI

Niall Gaffney
Director for Data Computing
Texas Advanced Computing Center (TACC)

# Sources of Information

- https://github.com/microsoft/generative-ai-for-beginners
- https://ollama.com/
- https://docs.openwebui.com/
- https://www.youtube.com/c/3blue1brown




1956

## Artificial Intelligence

the field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

1997

## Machine Learning

subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

2017

## Deep Learning

a machine learning technique in which layers of neural networks are used to process data and make decisions

2021

## Generative AI

Create new written, visual, and auditory content given prompts or existing data.

# Flow of Generative AI

- Create a computationally useful representation of what is being modeled ("Enbeddings")
  - Tokenization for text
  - Images use pixel and pixel location (movies add time)
- Create base model with Transformer to create a foundational model with weights assigned based on attention
- Use transformer to tune weights based on input data (tuning attention)
- Predict what the most likely output representation is
  - Sometimes, refine this based on context
- Translate into the expected output format (Text -> AI -> Image)

# Tokenizing

- What is a tokenizer?

# Tokenizing

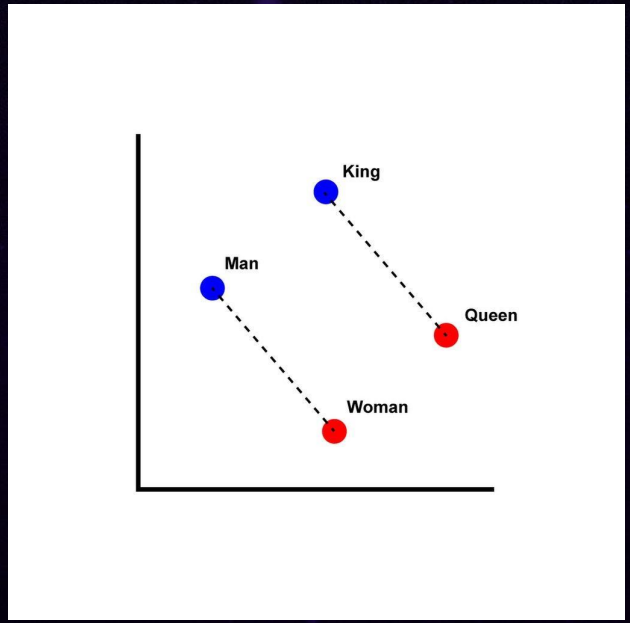
- What is a tokenizer?

# Tokenizing

- [2016, 422, 196, 22321, 982, 25]

# Where it started – Word2Vec in 2013

- Simple model. Go through a large corpus of sentences and assign a token to each word.
- Create a vector of associated words based on proximity in a sentence.
- Look at relationships in the vector-space between words



# Enter Transformer

- Transformer is composed of only encoders and decoders.
- All encoders have the same structure. Same for the decoders.
- The main building block is multi-head attention block, and the main technique is self-attention mechanism.

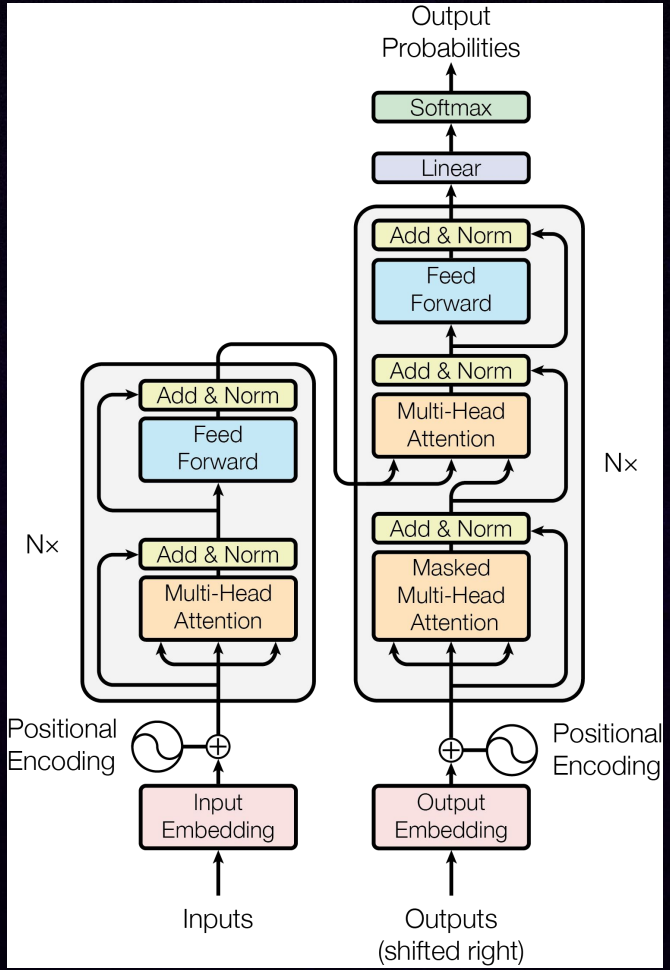



Image Credit

# Can I train my own model

- GPT4 cost an estimated $100,000,000 to train on a vast amount of data, expensive hardware, and power
- Can train a model on more limited data for more constrained results
- Other methods allow one to refocus a generalized foundational model to address specific needs

# Prompts

- Prompts are more than simply asking a question. They serve to help align the transformer to work with weights aligned to the goals of the query.
  - "Assuming the character of Albert Einstein and given what had been proven about General Relativity in 1950, explain the gravitational waves we expect to see from colliding black holes" vs "How do gravitational waves work"
  - You can make the prompt be the rules of a game
- Prompts only work if the foundational model has enough data about the information being prompted

# Fine Tuning

- Fine Tuning is taking an existing models weights and refining the weights of that model based on additional data
  - Tunes the model to work better in the context being tuned to
  - We are taking Llama 3 and fine tuning it based on the use guides, forums, problem tickets, and other data collected at HPC centers to make an AI to help answer questions about how to work on these systems that are not what you would get for how to do something on any computer
- Requires retraining as more data are produced

# Retrieval-Augmented Generation - RAG

- RAG is a step that takes place before prompting a model. Some source of data is used to create a tokenized prompt used to refine the model prior to the users prompt
  - RAG could use a live web site with current status and information to add value to the questions being asked without fine tuning with this live data
- Requires recomputing the tokens for the RAG data at the start of every session

# What models are there out there

- https://ollama.com/library
- https://huggingface.co/models

# Getting Started on evaluation

- Do I need a supercomputer to work with different Ais?
  - No.  Welcome to ollama and open-webui
  - Install ollama downloaded from https://ollama.com/
    - Start an ollama process with **`ollama serve`**
  - Install python 3.11 (not 12) and then **`pip install open-webUI`**
    - Start the server with **`open-webui serve`**
  - Connect to http://localhost:8080

# Examples

- How to run Tensorflow on ls6 at TACC
  - First ask how to run Tensorflow
  - Next ask on ls6 at TACC (delusional)
  - Next add the TACC ls6 user guide and query
- Creating a prompt to change behavior

# Thanks

- Questions
- Contact: ngaffney@tacc.utexas.edu
- Any Issues on TACC Systems
  - Open a ticket at https://tacc.utexas.edu/about/help/