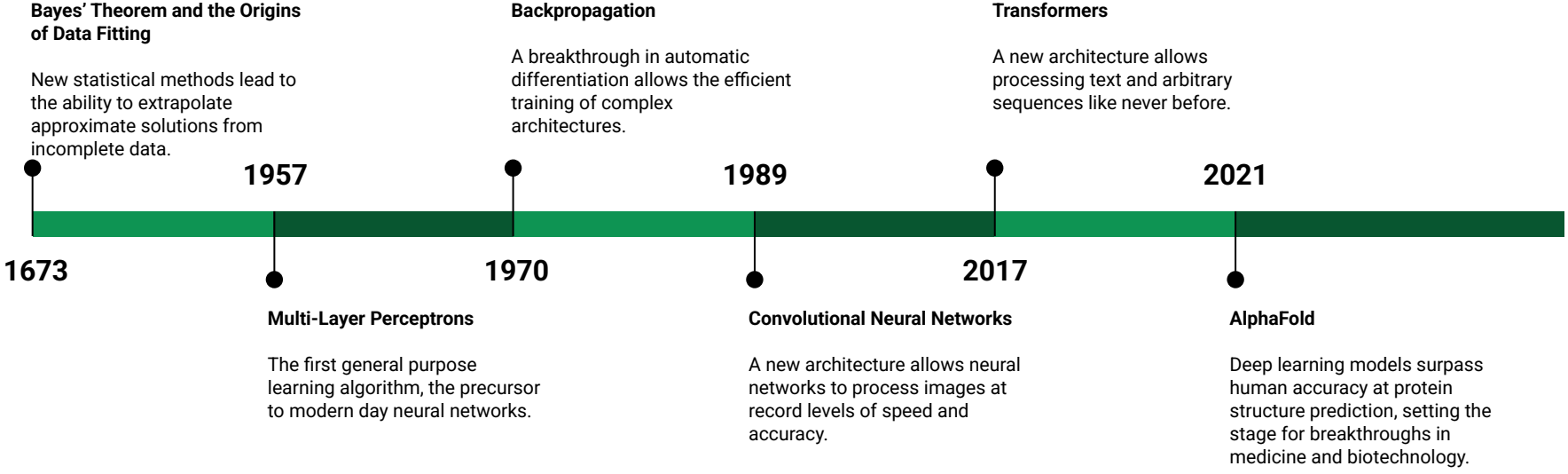
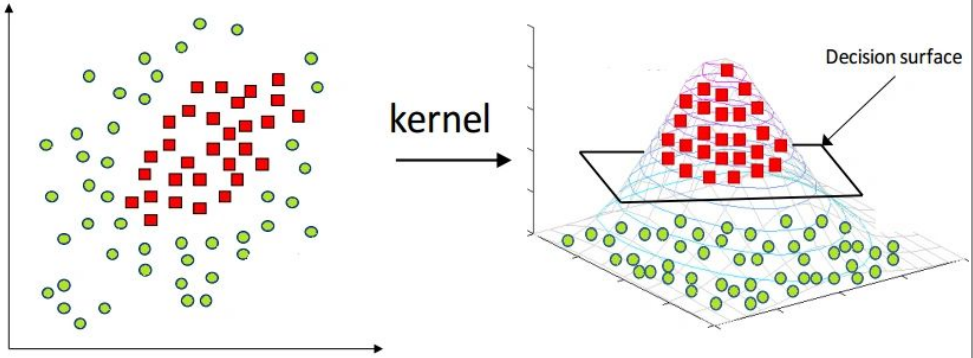
Introduction to Deep Learning & AI

The History of AI


Bayes' Theorem and the Origins of Data Fitting
New statistical methods lead to the ability to extrapolate approximate solutions from incomplete data.
Backpropagation
A breakthrough in automatic differentiation allows the efficient training of complex architectures.
Transformers
A new architecture allows processing text and arbitrary sequences like never before.
1673
1957
1970
1989
2017
2021
Multi-Layer Perceptrons
The first general purpose learning algorithm, the precursor to modern day neural networks.
Convolutional Neural Networks
A new architecture allows neural networks to process images at record levels of speed and accuracy.
AlphaFold
Deep learning models surpass human accuracy at protein structure prediction, setting the stage for breakthroughs in medicine and biotechnology.

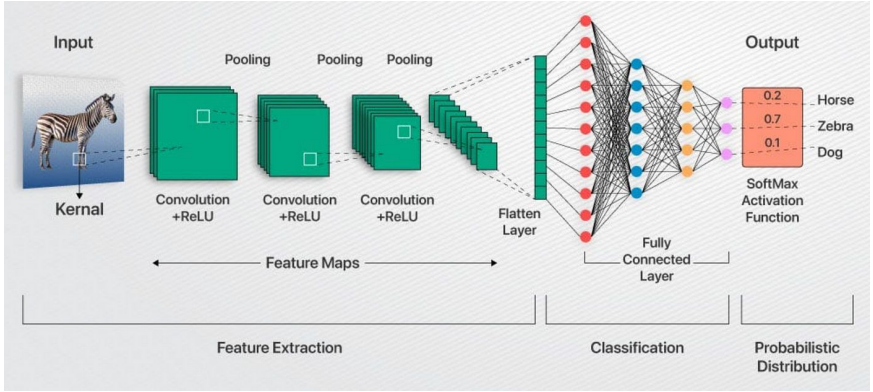
Machine Learning is Just Fancy Statistics

- Early machine learning models were effectively just applied statistics
- With simple, learnable, manipulations even complex data can be easily understood
- But the earliest model's struggled on anything other than basic, tabular data



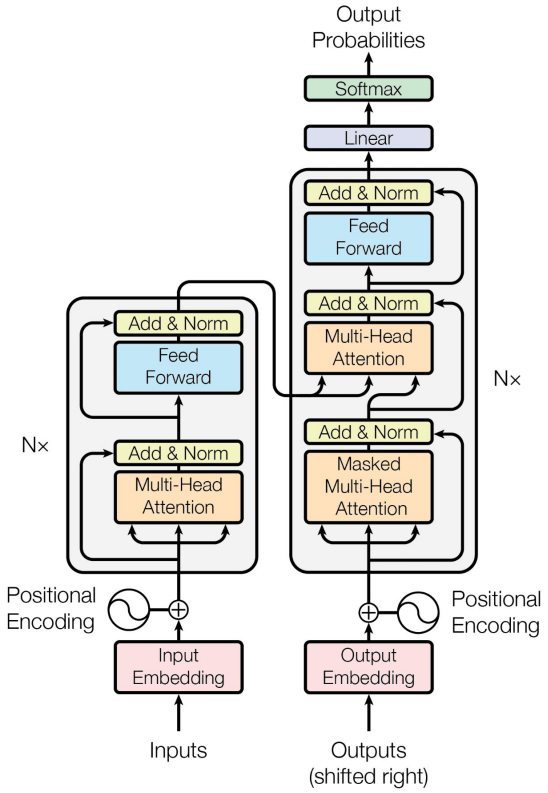
Convolutional Neural Networks

- Eventually, computing power caught up with ambition and the first computer vision models
- Faster, bigger, better, convolutional neural networks began pushing the boundaries of what was thought possible



Transformers

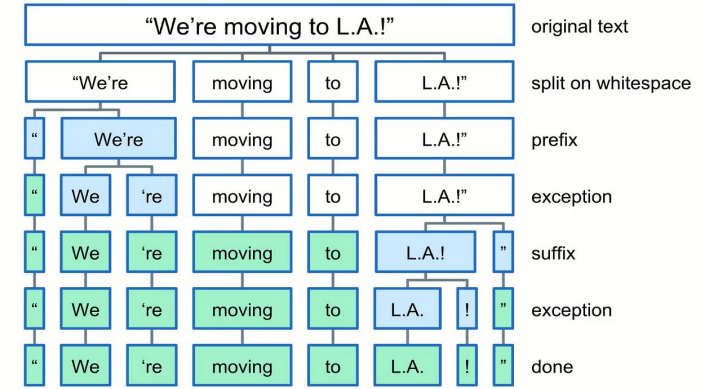
- After images, text was the next frontier to fall
- But it wasn't until 2017, with the invention of the transformer that this became possible



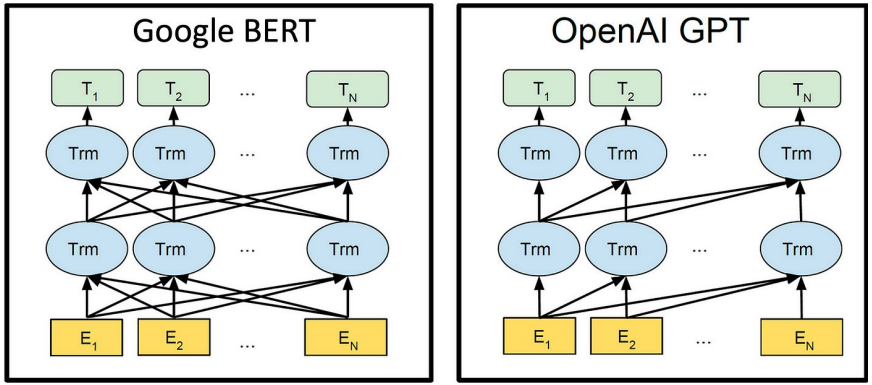
Large Language Models

How do LLMs Work?

- LLMs don’t read or write text the way humans do.
- First, the input text must be split into discrete “tokens”, each belonging to the model’s finite token set.
- Next, an generative text model (e.g. GPT) uses each token prefix of the input to predict the most probable next token.
- In contrast to autoregressive models suitable for text generation, bi-directional models can “see” the entire input sequence at every step, making them superior at tasks like sentiment analysis or text classification



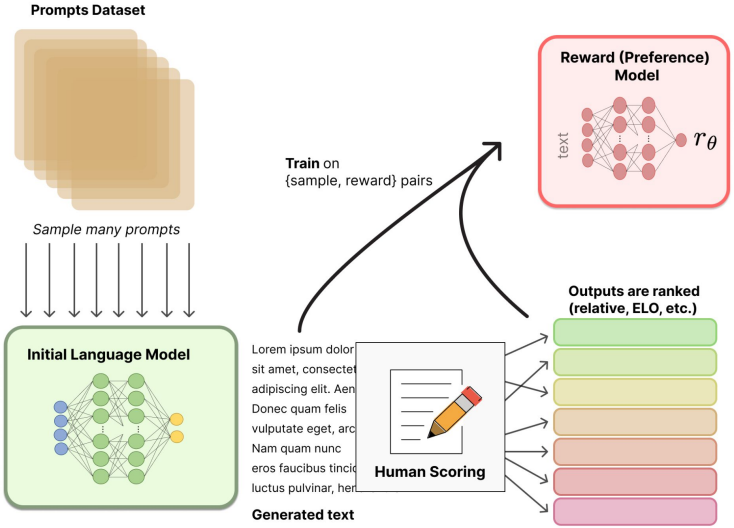
Tokenization



Bi-Directional vs. Auto-regressive

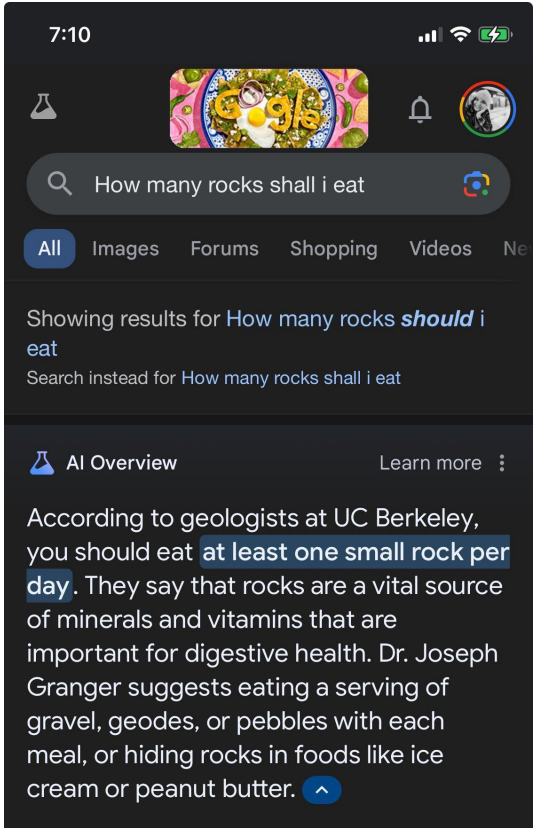
How to Train an AI assistant

1. Generative Pre-training
   - Given a large corpus of human-written text (10 trillion words for GPT4), train the model to predict each token given the previous n
   - Here, the AI learns grammar, syntax etc.
2. (Optional) Fine-tune on expert data
   - If your LLM is intended for a specific domain (e.g. code, medical data, etc.) you may want to restrict the later stages of your training to high quality text sources belonging to your domain.
3. Reinforcement Learning with Human Feedback (RLHF)
   - Train your LLM to respond to prompts in alignment with human quality rankings
4. (Optional) Retrieval Augmented Generation (RAG)



Hype vs. Reality

- A lot of people (particularly CEOs of large AI companies) make bold claims about LLMs
- But reality LLMs are a tool, far from able to replace humans
- Nothing an LLM says should be trusted without some extra process of verification
- But, that being said, they can be extremely useful in speeding up otherwise repetitive or tedious tasks



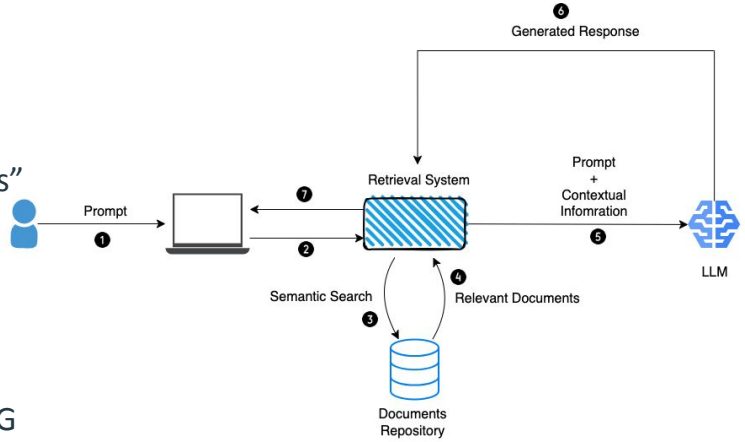
The (Not so Hidden) Cost

- LLMs are expensive. Extremely so
- Data gathering, human fine tuning and sheer compute costs meant that GPT4 cost around $100 million to train
- Even DeepSeek V3 which is famous for being cheap cost around $5 million
- So training a foundation model from scratch is probably not feasible for VeryConnect

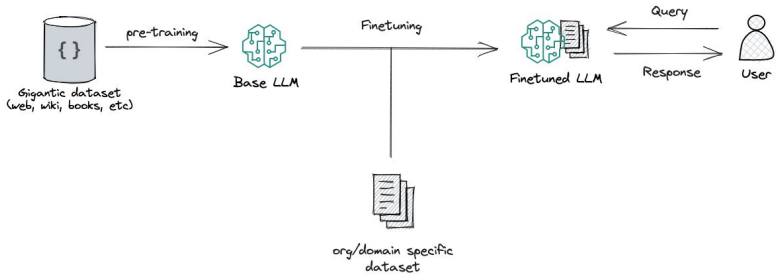
But Can They Work for You?

Retrieval Augmented Generation (RAG)

- How much information can an LLM remember? Can you trust an AI that's only trying to mimic human speech?
- Enter Retrieval Augmented Generation:
- Split a large repository of trusted information into "chunks"
- Train a separate retrieval model to find the chunks which will help the most in answering the user's prompt
- Append these chunks to the user's prompt before generating a response with your LLM
- Openai, deepmind etc. all offer cost-effective tools for RAG



Fine-Tuning



- A pretrained foundation LLM can be put through an extra step of post-training on bespoke domain specific data
  - This will enable the model to learn industry specific information and be better suited to more niche tasks
- A surprising amount of LLMs are open source
  - DeepSeek
  - Llama
- Although their model's aren't open source, OpenAI offers the ability to fine-tune and host models
- The cost can be as low as $3.00 per 100,000 tokens (roughly 75,000 words)
- The dataset does need to be reasonably large though, but not prohibitively so. A thousand or so examples should do

Prompt Engineering and AI Agents

- Prompt engineering is the process of crafting inputs (prompts) to guide AI models in generating desired outputs effectively
- AI agents have demonstrated a surprising ability known as *in context learning*. Typical AI models can only perform tasks on which they were specifically trained, but if an LLM's prompt is correctly crafted it can generalise previously learned tasks to provide quality solutions to ones unseen during training
- Zero-shot prediction: The prompt contains no input/output examples
- One/multi-shot prediction: One or more examples can enhance an agent's ability to solve an unseen task
- An AI agent is an LLM built into a software stack to automate routine tasks
    - E.g. customer support, data generation, document summarisation etc.
    - Existing tools include llamaindex, langchain etc.