- The first language model is ELIZA released in 1966 which was an NLP model
    - Had preprogrammed answers based on keywords
    - Had many holes in its logic
- RNNs didnt begin evolving until 1972
    - First program to really predict the next word in a sentence
- In 2017 google released their documentation on transformers
    - Led to chatGPT
    - Used to translate one language into another
- Gpt released in 2018
    - Had 117 million parameters
- BERT released in 2018
    - 340 million parameters
    - Bidirectional text processing
- Tokenization
    - Separating words into separate tokens, usually down by separating the prefix or suffix from the word.
    - What is the tallest building? Will be separated into: What is the tall est building,
- Embedding
    - When LLM turn tokens into embedding factors, each token turns into a vector (-760,760)
    - All values represent a position in vector space related to the token
- Vector Database
    - Each embedding vector is places somewhere
    - Tokens that are similar or related are closer together, this is how LLM's can predict the next word
- How Transformers work
    - You extract information from the vectors making a matrix using a method called multi-head attention
        - Is an algorithm that outputs a set of numbers telling the model how much each word and their order contribute to the overall sentence
        - We then convert the output matrix and a word will have the same value as the output
        - Input matrix -> Output Matrix -> Natural Language
- The metric used to determine the effectiveness of a model is the perplexity, it will test the closeness to two tokens, if they are close then it will give a good score
    - Also use RLHF
        - Reinforcement Learning Through Human Feedback
- Fine Tuning
    - Altering a pretrained model for a specific use
- Some problems of LLM
    - Hallucinations
        - When an LLM will be confidently wrong or completely make up information
    - Cost

        - LLm cost a lot of hardware to maintain and to train
- Retrieval Augmented Generation(RAG)
    - Give LLM's access to data its not trained on
- Context window
    - How much info you can give to a prompt to get an output

Softmax

- Transforms array to a probability matrix
- e^(each value)/(sum of e^(each value))

ChatGPT Terms:

- Prompt Templates
    - Templates for structuring input text given to an LLM
    - Mix of fixed input and dynamic variables
- Chains
    - Sequence of steps involving LLMs or functions, where outputs of one step become inputs to the next
- Agents
    - Dynamic systems that use LLMs to decide what to do next
- Memory
    - Allows the LLM to remember previous interactions across steps
- Indexes
    - Sturctures that store and organize data
- Output Parsers
    - Tools to structure and extract specific information from an LLM's raw response