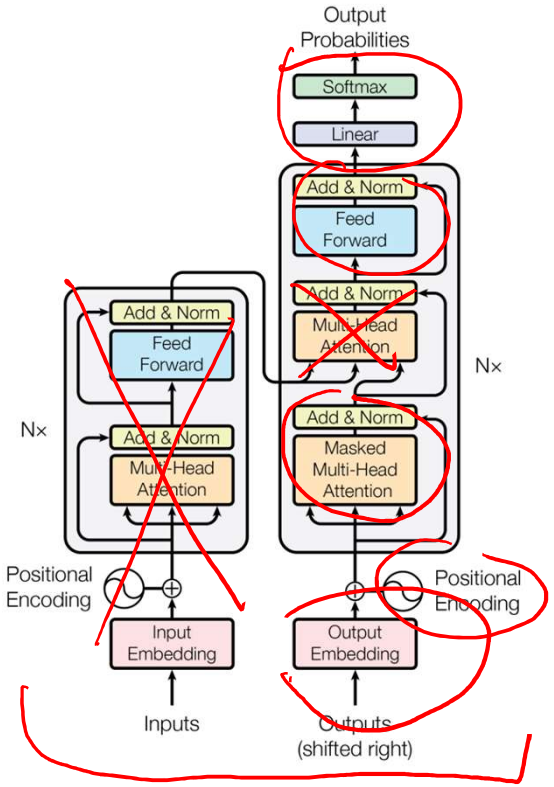
Outline

In this session, we'll discuss:

- Introduction to Large Language Models (LLMs)
- Training of LLMs
- Evaluation of LLMs
- Applications of LLMs
- Prompting

Large Language Models



- An LLM is a transformer neural network
  - Large: Billions of parameters
  - Language: Trained specifically for language tasks
  - Model: Transformer
- Usually, LLM is a reference to a generative decoder

Large Language Models

- GPT1 – 2018
- GPT2 – 2019
- GPT3 – 2020
- chatGPT – 2022
- PaLM, LLaMA, Claude, FLAN, Mixtral, Gemini, …
- https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Large Language Models

- GPT-3 was the first model that was truly spectacular.
  - 175Bn parameters
  - Trained on 300Bn tokens
- GPT-4 is even bigger
  - It took ~3 months to train
- In 2022, OpenAI bought ~30,000 NVIDIA GPUs.
  - Tens to hundreds of millions of dollars

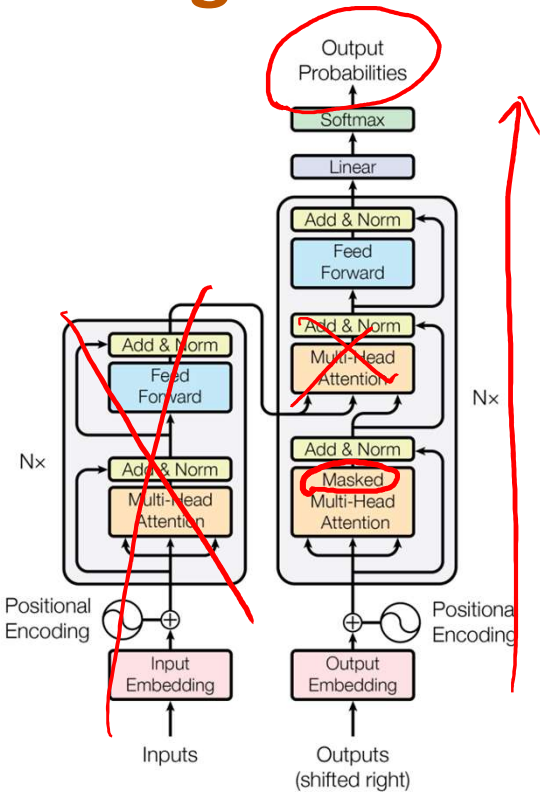
LLM Expectations

- Articulate
- General Knowledge
- "Creativity"

LLM Caution

- Specific domain knowledge
- Bias
- Hallucination

Training



- Web-scale data aggregation
  - Legal issues?
- Separation
- Tokenization (probabilistic)
  - Dictionary
- Training
  - My name is Dan and I work at UT.

Training

- Fine tuning
- Reinforcement learning human-human feedback
- Textbooks are all you need

Evaluation

- GLUE
  - Collection of 9 language tasks
- BLEU
  - Language translation
- ROGUE
  - Summarization and translation
- HellaSwag
  - Common sense inference

Multi-Modal LLM

- Use an LLM to generate an image

- Use an LLM to describe an image

Prompts

- A prompt is the input text you feed to an LLM
  - How much wood could a
  - How many meters are in a mile?
- The LLM then predicts the next word, repeatedly, until the predicted next word is [EOS]
- Some prompts will get better responses than others!
  - Why?

Prompt Engineering

- Can we write a prompt that is more likely to get a good response?
  - YES: Prompt engineering!

- Can we guarantee a perfect answer by engineering a prompt?
  - NO!

Prompt Engineering

- Control model behavior
- Get constrained outputs
- Higher output quality
- Automate operations

Prompt Engineering Limitations

- High sensitivity
- Common sense
- Exception handling
- Debugging

Prompt Engineering

- Template
  - Translate this sentence from English to Spanish:

- Fill in the blank
  - The first person to walk on the moon was _____

- Multiple choice
  - Which of the 3 options below is correct?

Prompt Engineering

- Instructional
  - Write a neutral-tone sales pitch in 300 words for a pair of socks
- Iterative
  - Start with a broad prompt and narrow it based on LLM's output
- Ethically aware
  - LLMs may avoid answering certain questions

Applications

- Healthcare
  - DocBot
  - Personalized treatment plans
- Retail
  - Virtual shopping assistant
- Tech
  - Code generation
  - Marketing campaign management

Applications

- Retrieval Augmented Generation (RAG)
  - Submit a query
  - Search a database for relevant entries
  - Use found entries in prompt to LLM

RAG Systems

- Take a large database of documents.
- Embed every entry in the database.
  - Sentence-transformers
- Compare the query's embedding to each embedding in the database.
  - Return top k results
- Engineer a prompt to LLM to include questions and relevant results.
  - Answer the following question using the context below.

RAG Software

- Pinecone/FAISS/LangChain
- C3 AI
- Westlaw Precision

Ethics

- Ethical usage of AI
- Ethical training of AI
- Guardrails

Conclusion

Here's a brief recap:

- LLMs are used to generate text.
- They are trained using a large corpus of text.
- There are MANY applications of LLMs in practice.
- Prompt engineering is important to get appropriate responses.