Introduction to Large Language Models

What is a Large Language Model?

A large language model (LLM) is a type of artificial intelligence (AI) that excels at understanding and generating human-like text.

They are trained on massive datasets of text and code, enabling them to perform a wide range of linguistic tasks.

Two Components of LLMs: Data + Neural Network

Training Data

LLMs learn from vast amounts of data.

This data is unstructured and usually captured from the internet.

This data helps the model develop an understanding of language structure, grammar, and context.

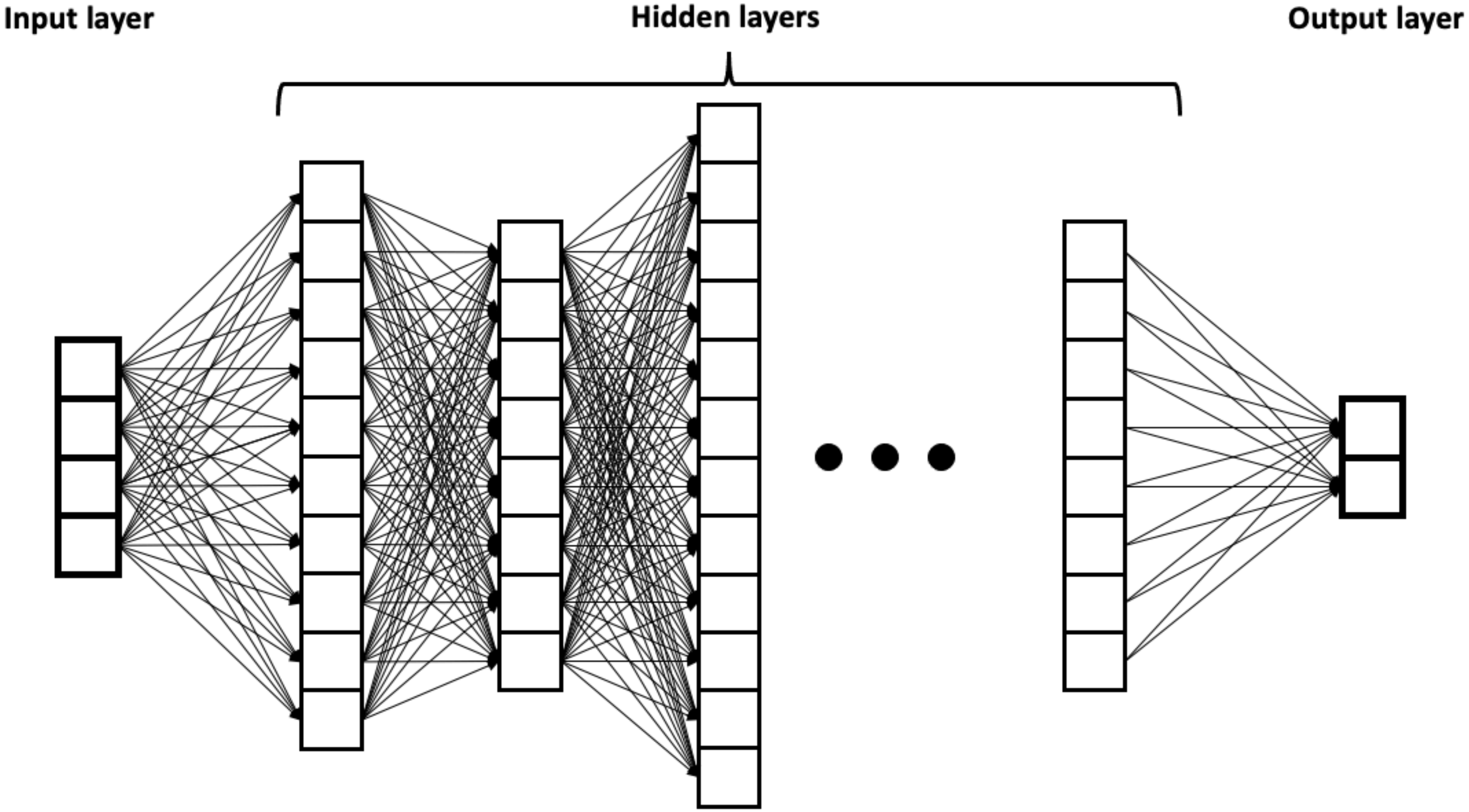
The Pile Dataset

One popular dataset called The Pile contains over 100 billion words (825GB) from publicly available sources like Wikipedia, books, and online articles.

This massive dataset provides the model with a broad understanding of language patterns and nuances.

The Pile also introduces all of the inconsistencies, deviant behavior, and toxicity you find on the internet.

Deep Learning AKA Neural Networks



How Deep Learning Works

Neural networks are the core component of LLMs. They process data in layers, much like a tree with branches. Each layer learns to extract more complex features from the previous layer.

The weights are adjusted during training through a process called **backpropagation.** This helps the model learn and improve its accuracy over time.

Backpropagation: The Learning Process

- It starts with an initial guess for the output of the neural network.
- Then, it calculates the difference between this guess and the actual output.
- This error is propagated backward through the layers, adjusting the weights until the output matches the desired result.

The Training Process

Pre-training

This pre-training masks words in text and the model learns to predict that word. After hundreds of training cycles the model becomes very good at predicting the next word using it's training data.

However this by itself is not very useful. These models ramble on with no grounding or direction. What is needed is task specific fine tuning.

End Use Case Fine Tune

Once trained on a massive dataset, LLMs can be fine-tuned for specific tasks.

— Translation

— Summarization

— Writing different kinds of creative content

— Chatbots

Fine Tuning for Services

For services like Gemini, Claude, or ChatGPT, models are fine-tuned on question-answer pairs or longer conversations.

- This makes the model behave in a specific way.
- Fine tuning can also steer the model with specific guidelines:
    - Harm reduction
    - Helpfulness
    - More

Example: Scientific Articles

A model trained on a large corpus of scientific articles might excel at:

- Summarizing research papers
- Generating technical documentation

Predictive Text Generation

LLMs can predict what comes next in a sequence of words based on their training.

This ability allows them to generate coherent and grammatically correct text.

For example, if you ask an LLM "What is the capital of France?" it will likely respond with "Paris."

Key Issues With LLMs

LLMs are not Reasoning

It is important to remember that these LLMs are not reasoning.

They are simply predicting the statistically likely next token.

Examples of LLM Capabilities

LLMs can be trained to do many things. Large "foundation models" (GPT4, Gemini, Claude, etc.) are capable of many tasks. However it is also possible to have a smaller model trained to do specific tasks well.

- Text Summarization
- Translation
- Creative Writing
- Question Answering
- Chatbots

Hallucination

LLMs have a tendency to invent things.
- This is an ongoing issue.
- Cleaning pre-training data and using better fine-tuning data can help reduce the problem.

In short, be cautious about what LLM outputs.

Are LLMs Just Compression? Search?

Compression?
- They can be seen as a unique type of data compression.
- LLMs remember and retrieve information in a user-friendly format.

Search?
- LLMs are convenient search engines that sometimes fabricate information.

Explainability Challenges

- LLMs are often called "black boxes" because we can see the input and output, but not the internal processes that lead to a response.
- This opacity makes it difficult to understand *why* an LLM generates a specific output.

Impact on Trust and Applications

Lack of explainability impacts trust in LLMs:
- It's hard to assess reasoning behind outputs.
- Identifying potential biases becomes challenging.
- Difficult to use LLMs in critical applications where transparency is crucial.

A Brief Timeline of LLMs

2010-2017: Foundations and Early Innovations

- **2013**: Word2Vec was introduced by Tomas Mikolov and his team at Google. This method learned word embeddings from raw text, improving natural language processing (NLP) tasks significantly.
- **2015**: The Attention Mechanism was introduced, enhancing neural machine translation by allowing models to focus on specific parts of input sequences.
- **2017**: The seminal paper "Attention Is All You Need" by Vaswani et al. introduced the Transformer architecture. This architecture became foundational for modern LLMs and allowed for more efficient handling of sequences compared to previous RNNs and LSTMs.

2018: Emergence of Key Models

June 2018: OpenAI released GPT (Generative Pre-trained Transformer), showcasing the capabilities of unsupervised learning for text generation.

October 2018: Google introduced BERT (Bidirectional Encoder Representations from Transformers), which improved the understanding of context in language tasks through bidirectional training.

2019-2020: Expansion and Scaling

- **2019**: OpenAI released *GPT-2*, an improved version of GPT with 1.5 billion parameters, demonstrating advanced text generation capabilities.
- **2020**: OpenAI launched *GPT-3*, a massive model with 175 billion parameters, setting a new standard for LLMs in generating human-like text and performing various NLP tasks.

2021-2023: Specialization and Multimodality

- **2021**: Google introduced *LaMDA* for conversational applications, and OpenAI released *DALL·E*, a multimodal model capable of generating images from textual descriptions.
- **2022**: Google released *PaLM*, a large model with 540 billion parameters, continuing the trend of scaling up LLMs.
- **November 2022**: OpenAI launched *ChatGPT*, based on the GPT-3.5 model, which gained widespread attention for its conversational abilities.

2023: Continued Advancements

- **March 2023**: OpenAI released *GPT-4*, a more advanced and versatile model than its predecessors, with improvements in understanding and generating text across various contexts.
- **Today**: Claude 3.5, Gemini 1.5 Pro, Llama 3.1 405B are all competing for the top spot.

Questions?