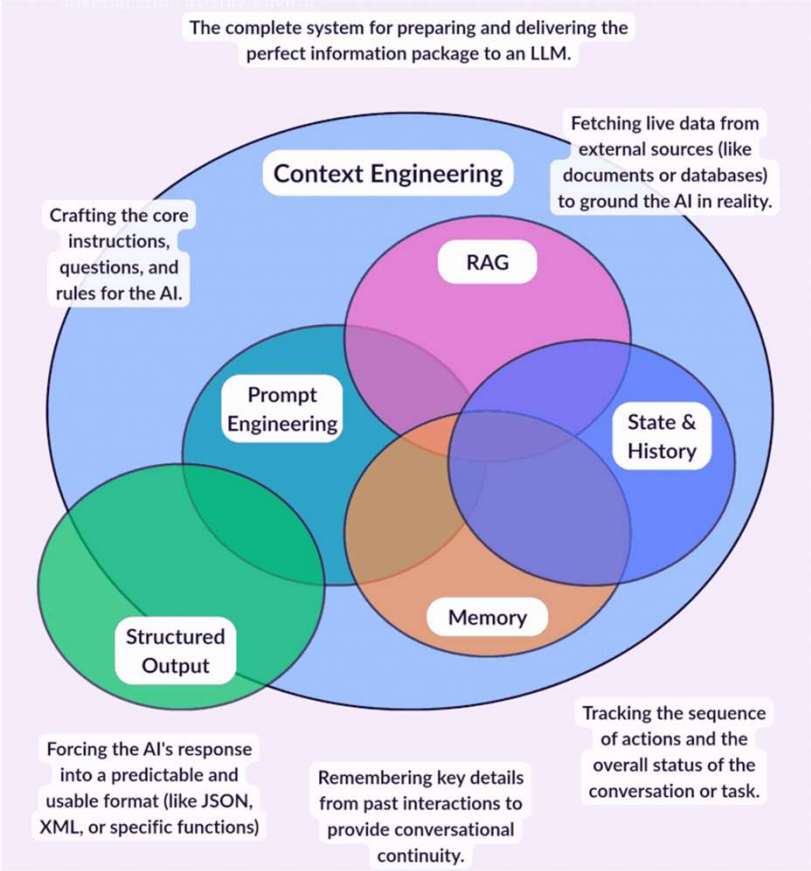
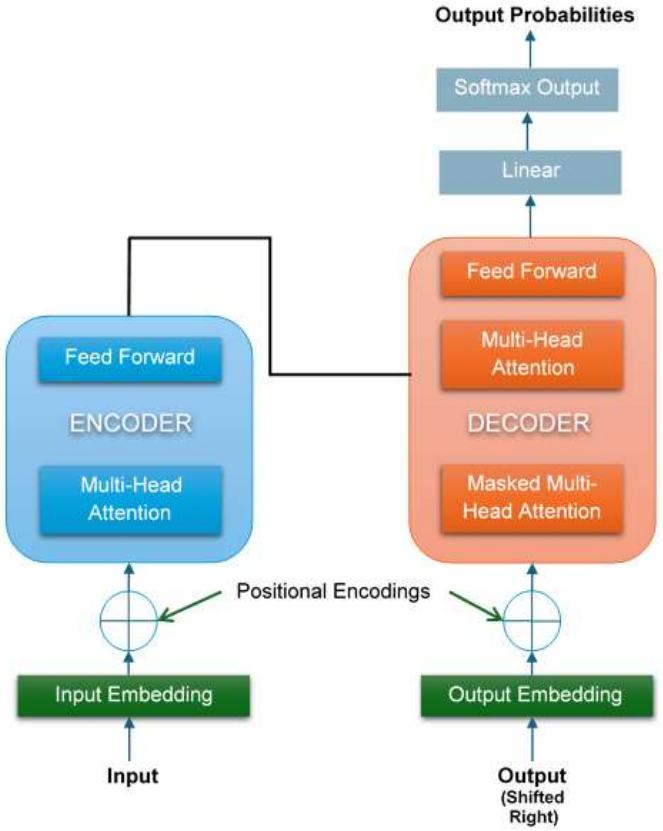
The Backbone of Agentic AI

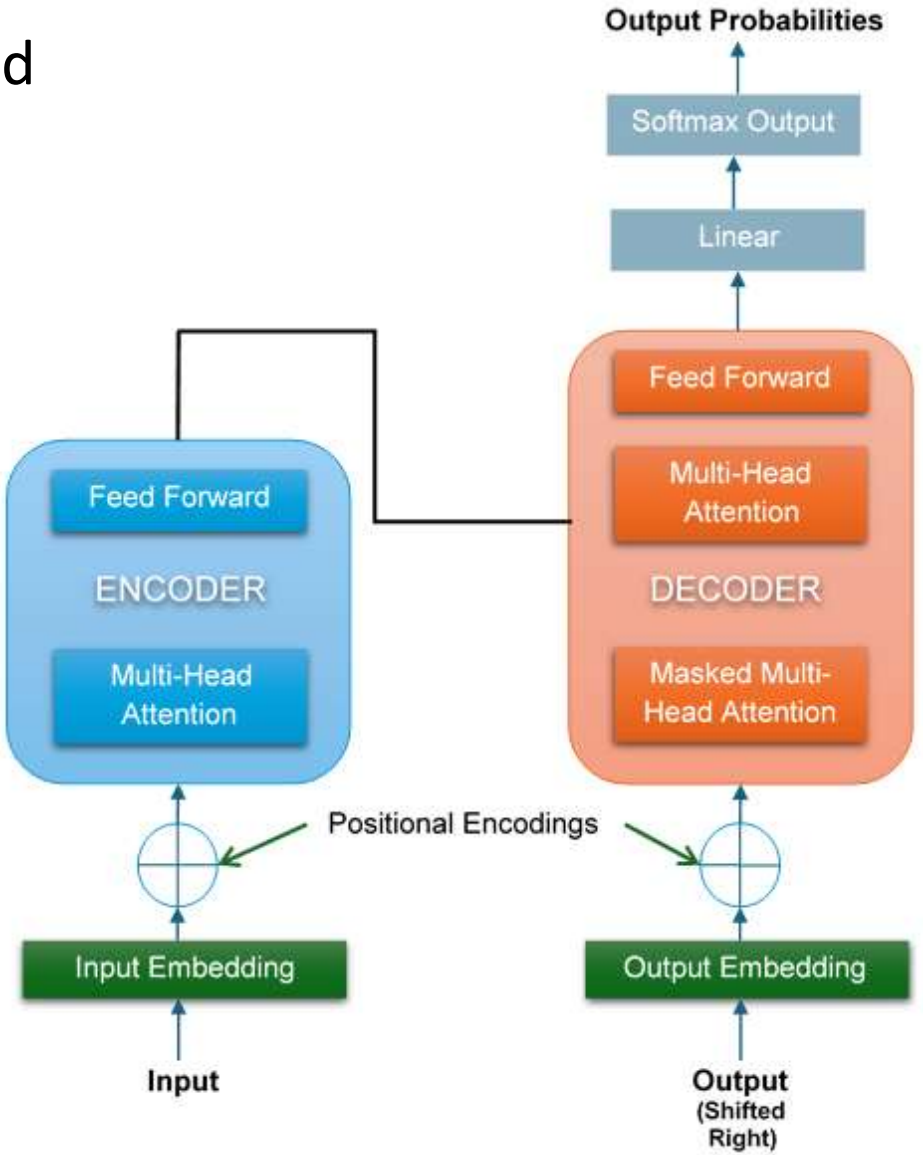


Why Transformers?



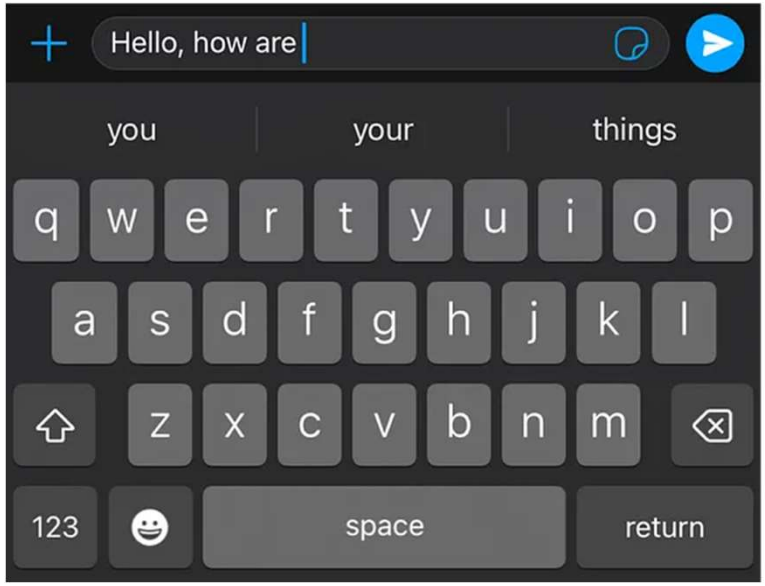
What is Transformer?

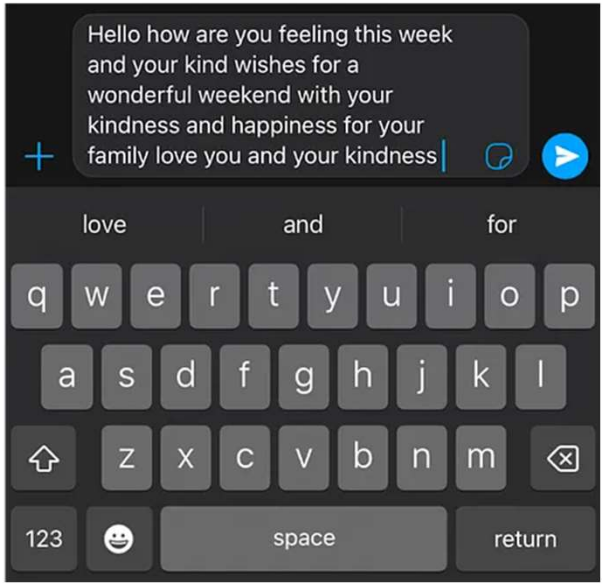
- Introduced in paper Attention is All You Need
- Utilizes a "self-attention" mechanism
  - To process sequential input data
- A deep learning model architecture
- Primarily used for NLP and computer vision
- Process inputs in parallel



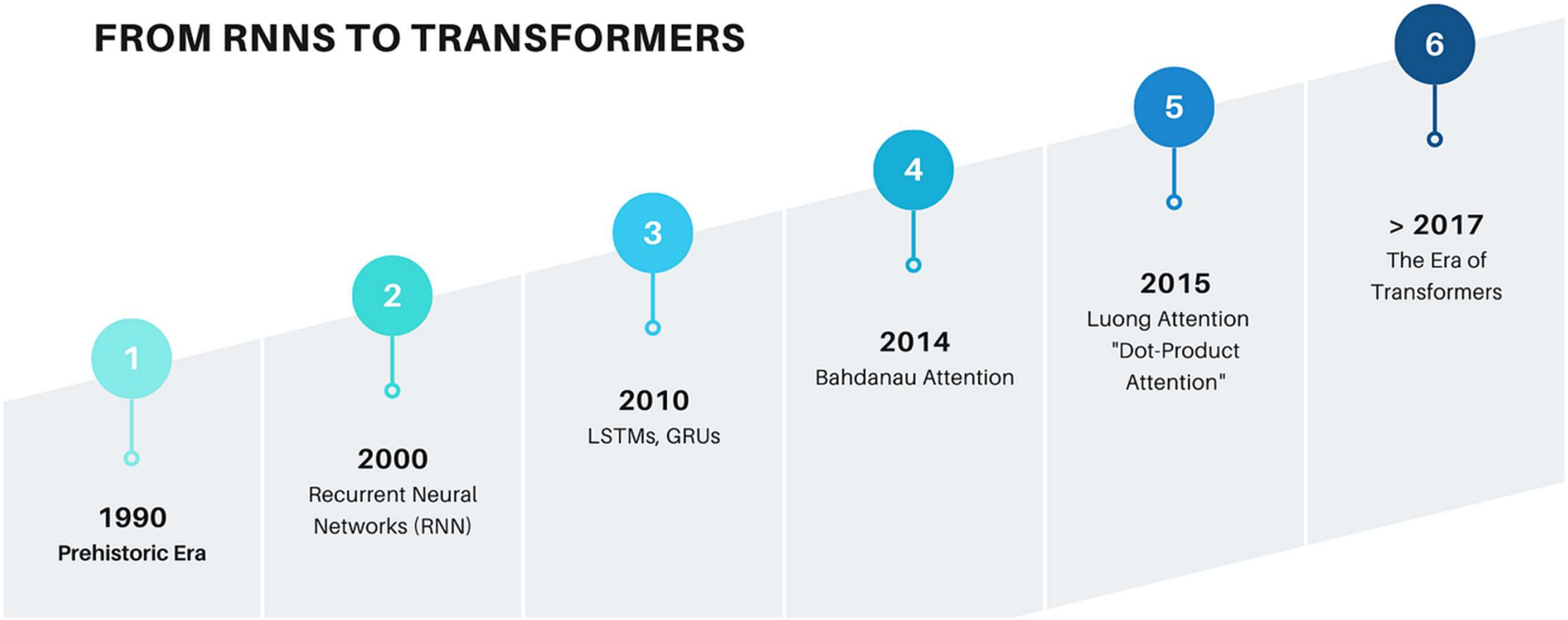
Understanding Transformer Models

- Can be used to:
  - Write stories, Essays, poems
  - Answer questions
  - Translate between languages
  - Chat with humans
- The phone can suggest the next word to use in a text message



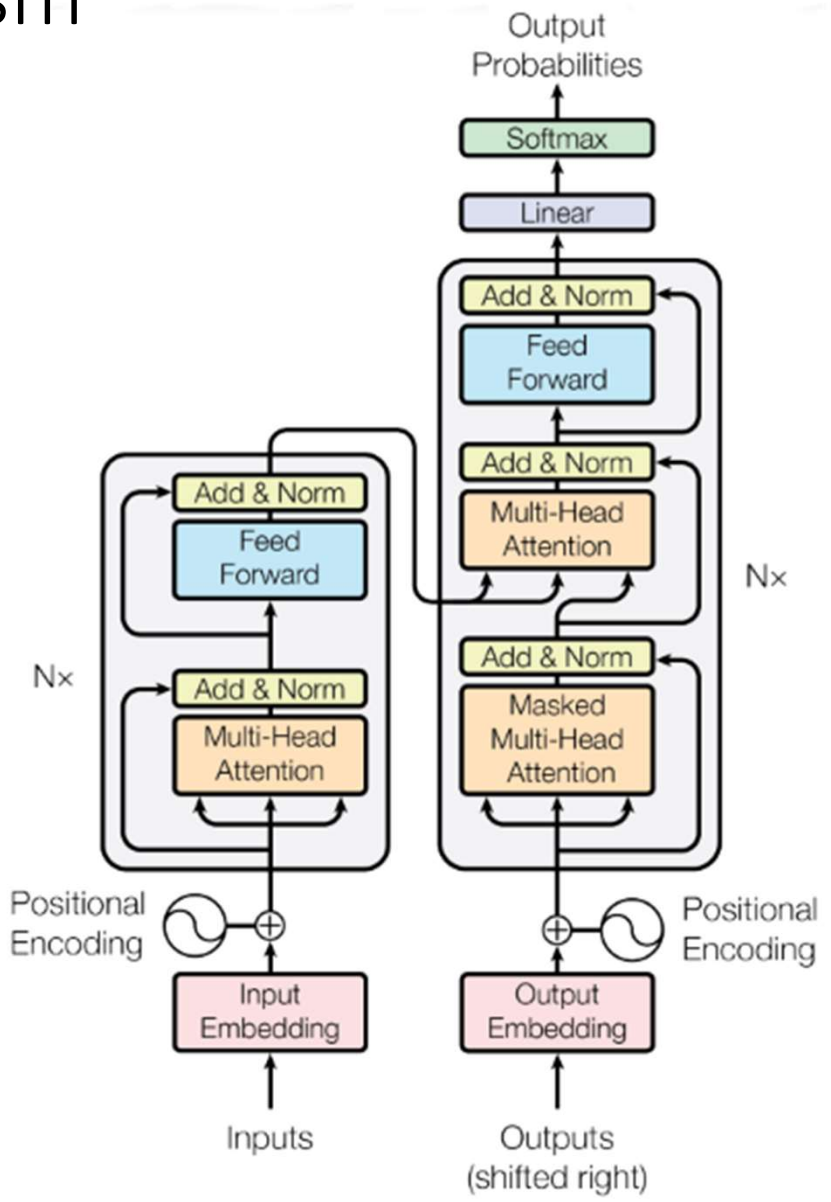


Evolution from RNNs to LSTMs to Transformers



Introduction to Attention Mechanism

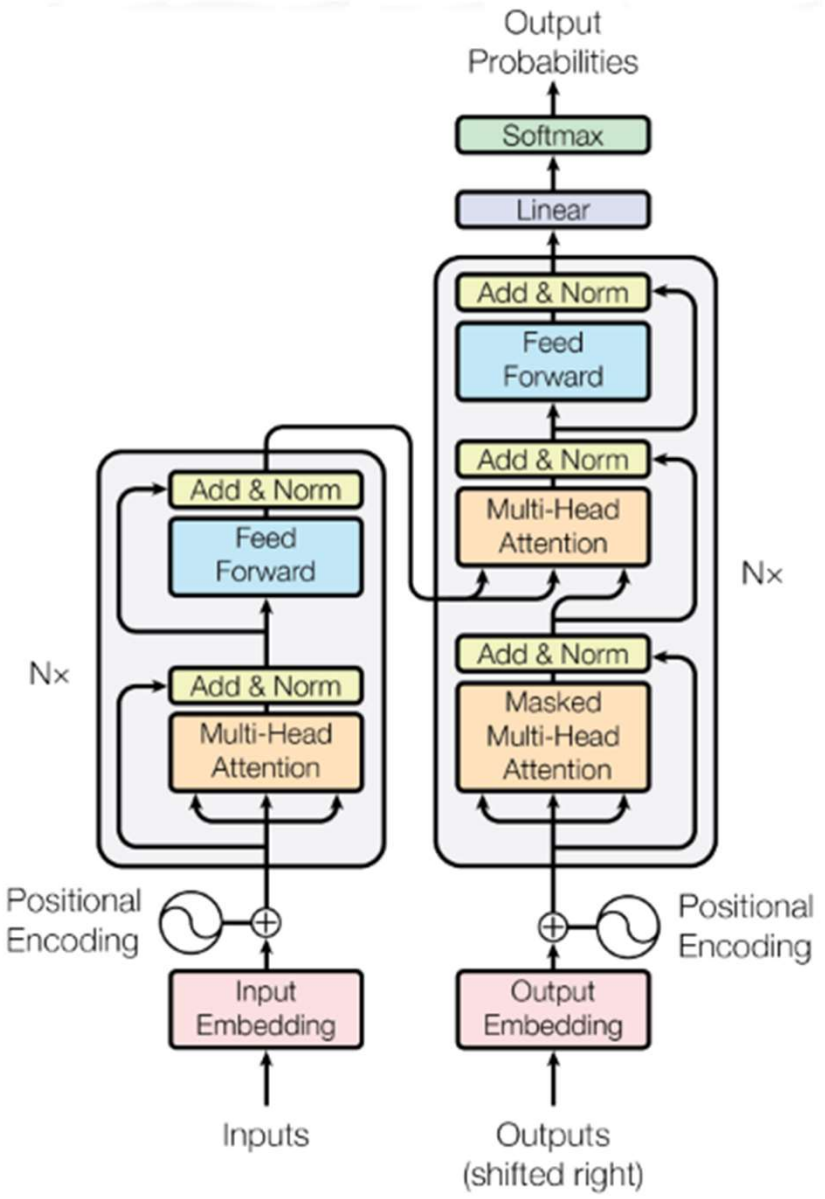
- Allows a neural network to weigh the importance of different parts of the input data
- It dynamically calculates "attention scores" for every input element relative to another.



Transformers Model Architecture

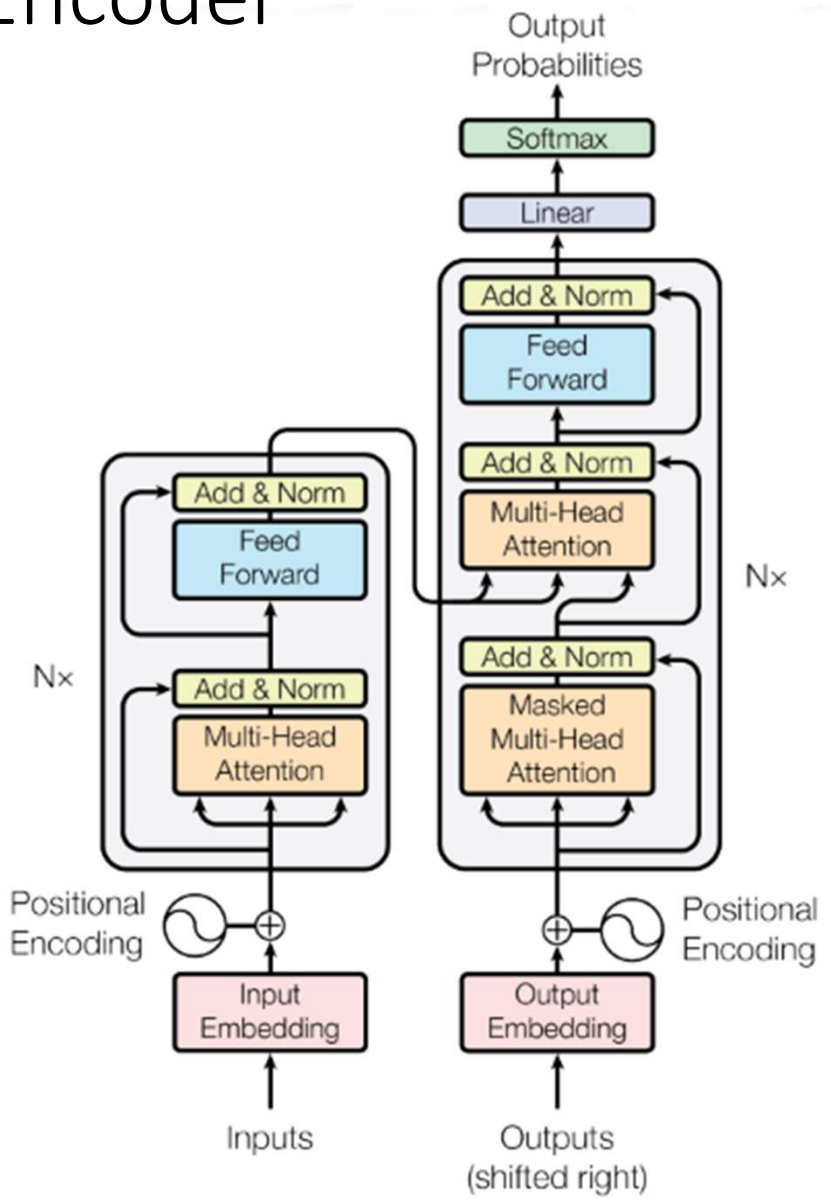
- Encoder
- Decoder
- Attention
- Feed-Forward Networks (FFN)
- Layer Normalisation
- Positional Encoding

- Follows an encoder-decoder structure
  - An input vector x is fed into an encoder
  - Resulting in a mapped representation z
  - With z, the decoder generates an output y



Transformers Model Architecture: Encoder

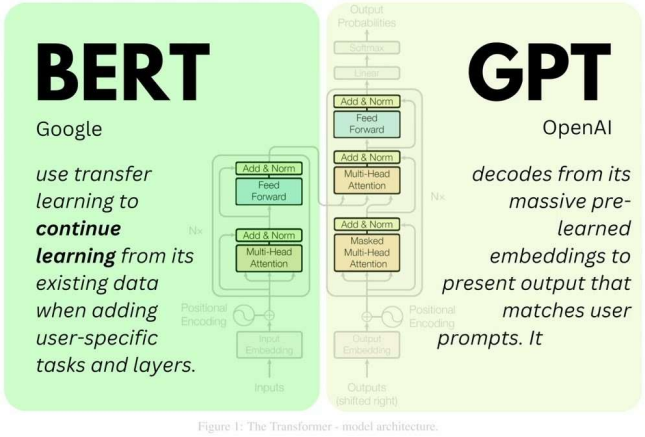
- N = multiple identical stacked layers



Transformers Model Architecture: Encoder

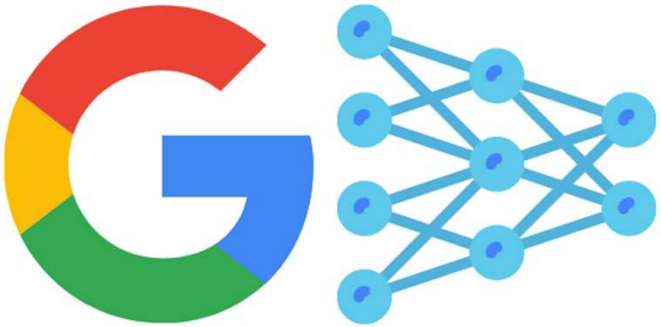
Overview of BERT, GPT, and LLMs in AI agent development

- LLMs act as the central reasoning engine for modern AI agents
  - Enabling them to understand, plan, and act.
- BERT:
  - The "Understanding" Engine
- GPT
  - The "Reasoning & Response" Engine
- The Modern Agentic Loop (LLMs Today)
  - Combine both strengths in a Reason → Act → Observe loop



BERT

- What it does
  - Uses bidirectional context to deeply understand the user's command.
- Role in Agents
  - Powers intent recognition and entity extraction
    - i.e., figuring out precisely what the user wants



GPT

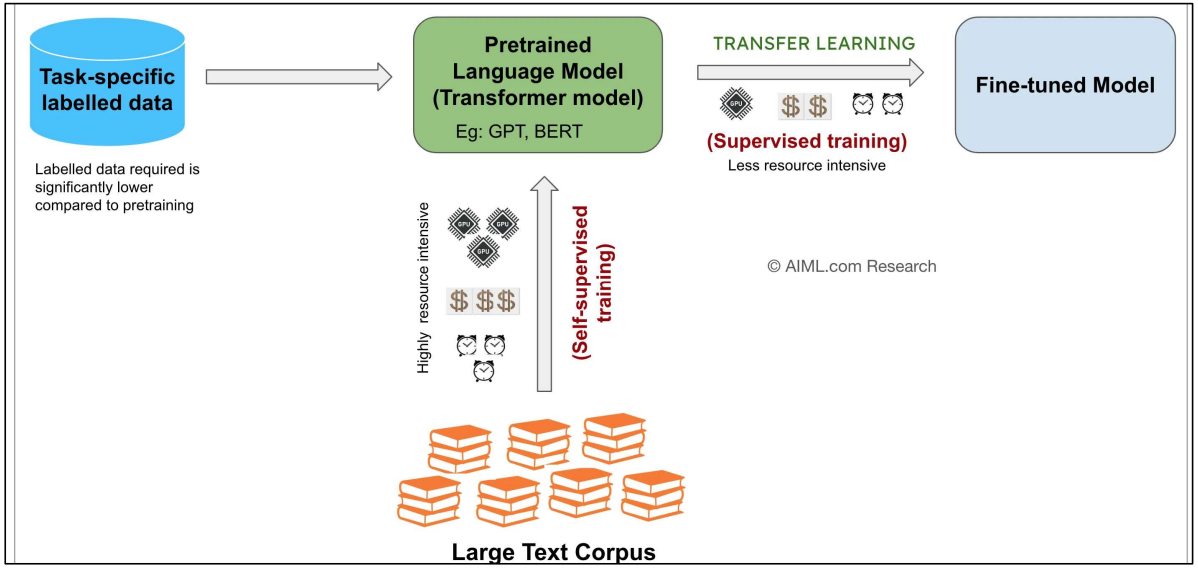
- What it does
  - Uses auto-regressive generation to
    - Create plans
    - Reason
    - Produce text
- Role in Agents
  - Enables
    - Planning (chain-of-thought)
    - Tool use (calling APIs)
    - Generating natural responses

LLMs

- The Process:
  - Modern LLMs combine both strengths in a Reason → Act → Observe loop
- Reason:
  - The LLM creates a multi-step plan
- Act:
  - It executes a step by calling a tool (e.g., search, calculator)
- Observe:
  - It analyzes the result and repeats the loop

- The Impact
  - This transforms the LLM from a chatbot into a dynamic problem-solver that can accomplish complex tasks.

Pre-trained Models and Fine-tuning



Hugging Face

Open-source platform for machine learning

- Often called "The GitHub for AI,"
- Provides the tools and infrastructure to
  - Build, train, and deploy state-of-the-art models

Provides the common ground for the entire AI community to

- Share
- Collaborate
- Innovate

Core Libraries: The Developer's Toolkit

Hugging Face: Core Libraries

Transformers

- Provides simple, unified access to thousands of state-of-the-art models

Datasets

- Standardizes the process of loading and preparing data

Accelerate

- Simplifies scaling model training across any hardware (multi-GPU, TPU)

Tokenizers

- Converts text into numerical tokens that a model can process

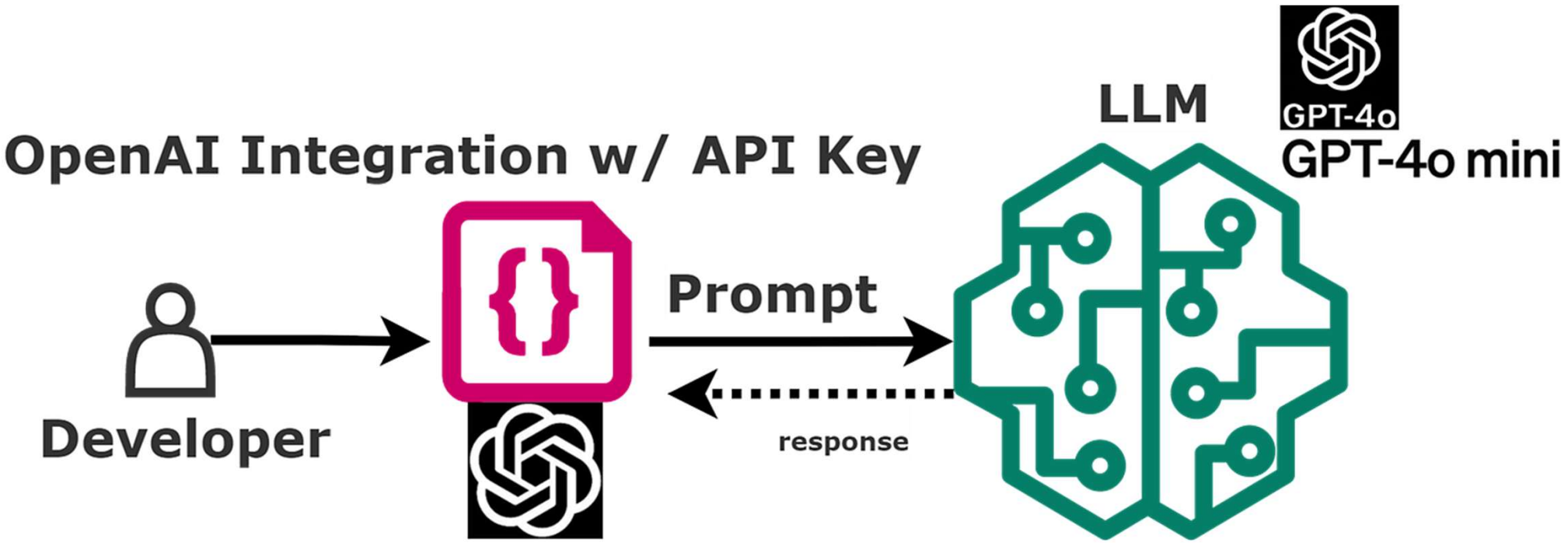
OpenAI APIs

Provides developers with programmatic access to OpenAI

Key Models

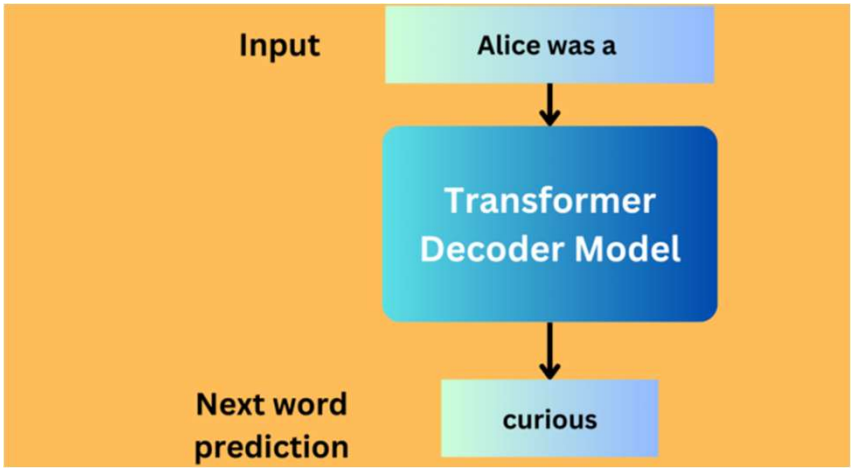
- GPT Models
  - Text understanding
  - Reasoning
  - Generation
- DALL-E Models
  - For generating unique images
- Whisper
  - Speech-to-text model
- Embedding Models
  - Convert text into a numerical vector representation

How Developers Interact: The Chat Completion API



Using transformers for text generation and classification

- For Text Classification (Understanding)
  - The Goal
    - To assign a label to a piece of text
      - (e.g., sentiment, topic, intent).
  - Analogy:
    - Reading a full product review to
      - Give it a 1-to-5-star rating

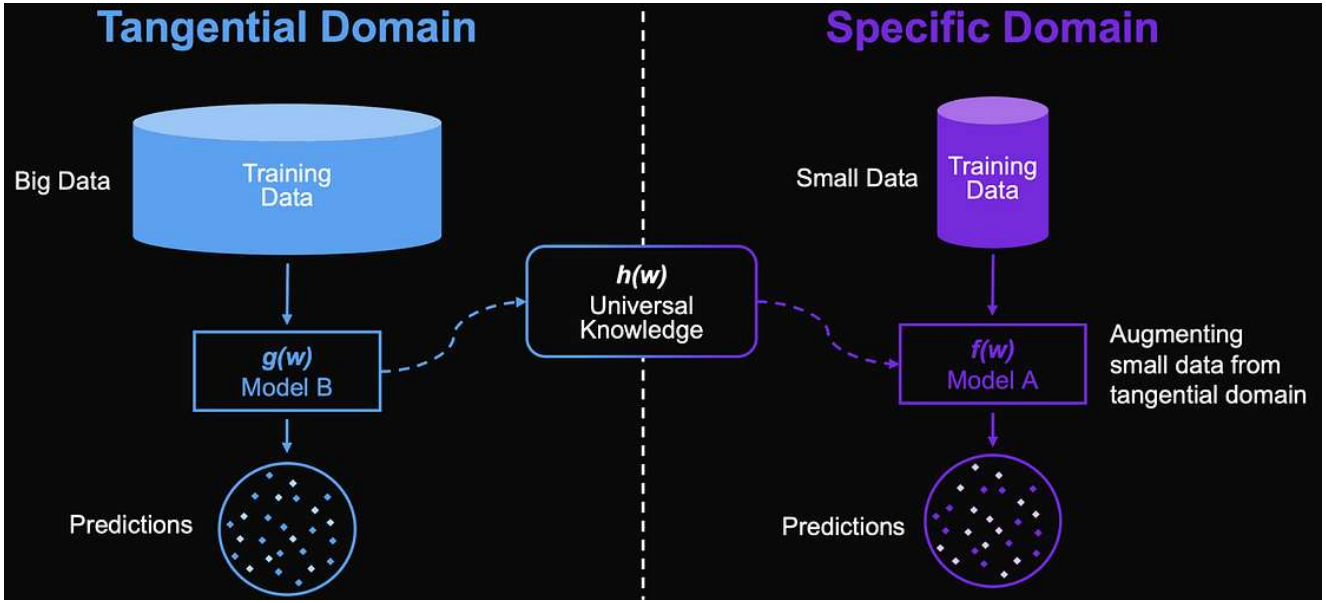


- For Text Generation (Creating)
  - The Goal
    - To create new, coherent text that follows a prompt
  - Analogy
    - Finishing a sentence by guessing the most logical next word, over and over

Transfer Learning & Domain Adaptation

- Don't train models from scratch.
- Leverage existing knowledge and adapt it

- The process of fine-tuning the pre-trained model on a smaller, specific dataset that is relevant to your unique "domain.“



Hands-on

- Using OpenAI's GPT API for text generation

- Fine-tuning a BERT model for sentiment analysis on real-world datasets


OCI
Generative AI

Introduction to OCI Generative AI Service

- Enterprise-focused
- Fully managed
- Providing secure access to leading AI models
- Designed for custom business needs
- Deep integration with private data
- Enables text generation, summarization, embeddings, and chat capabilities

Hands-on

- Prompt Engineering with OCI Generative AI Models
- Working with OCI GenAI Playground

Thanks