Generative AI Workloads

A Deep Dive into the Inner Workings of Autoregressive LLMs and How They Become AI Assistants

Agenda

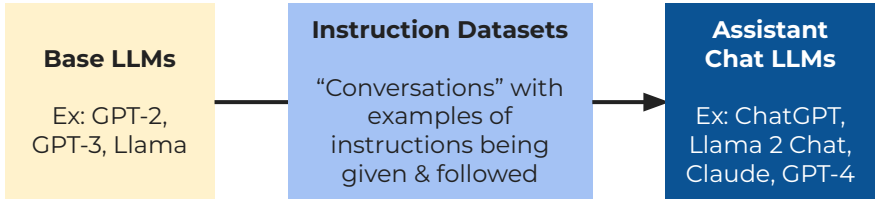
- Large Language Models - Base vs Chat Models
- LLMs - The Training Procedure
- The LLM Inference Mechanism
- The Evolution of LLMs to Chat Assistants
- OpenAI’s Assistant Training Recipe

Large Language Models (LLMs)
Base Models vs Chat Models

- **LLMs are Deep-Learning based AI models built on the Transformer architecture**. They are trained on a humongous volume of text data to achieve **high-quality text understanding & text generation.**
- The training regimen of LLMs usually consists of a simple objective such as **predicting the missing (masked) word in large text corpuses** - just the process of learning how to predict this missing word very well has been shown to imbibe an impressive understanding of grammatical rules, good sentence construction, reasoning power and real-world knowledge.
- There are two variants of Large Language Models that we need to differentiate:

  1. Base LLMs are purely trained on missing word prediction on generic internet text corpuses

  2. Assistant-style Instruction-tuned Chat LLMs however, have additional training steps that involve missing word prediction on curated high-quality "instruction datasets"

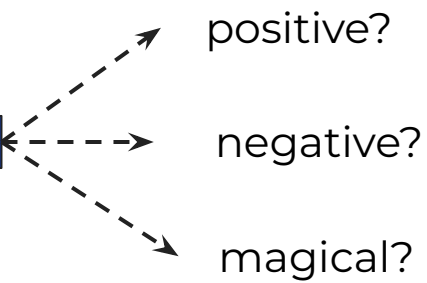


Large Language Models (LLMs)

*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*

Masked Sample

The movie was awesome. Overall, the experience was 

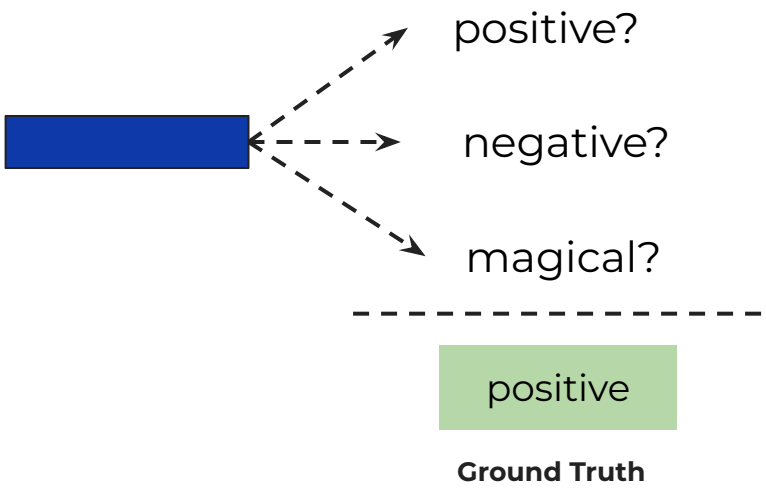


Large Language Models (LLMs)

*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*

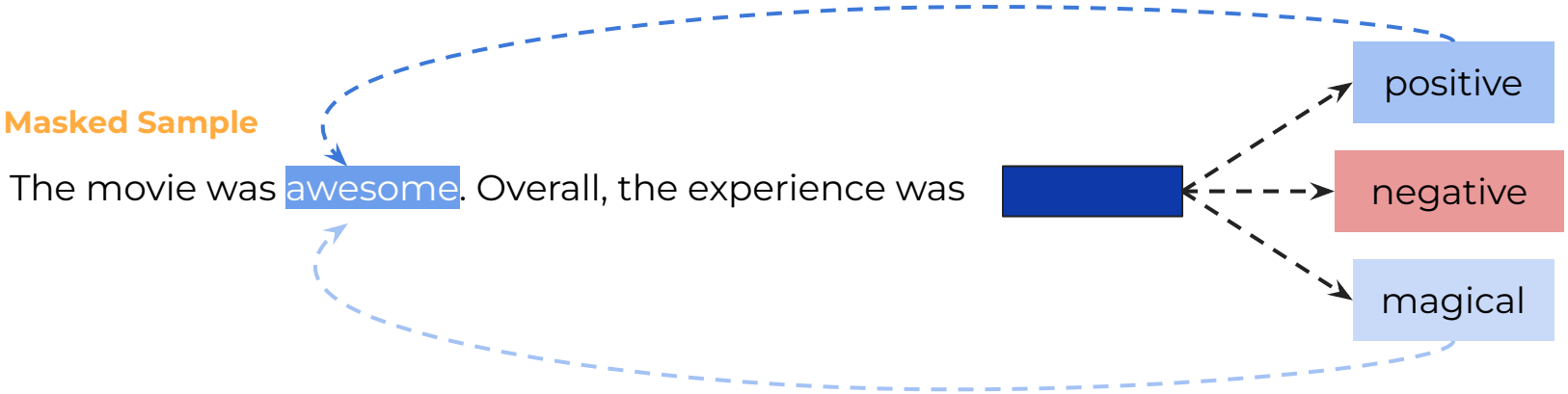
Masked Sample

The movie was awesome. Overall, the experience was [redacted]



Large Language Models (LLMs)

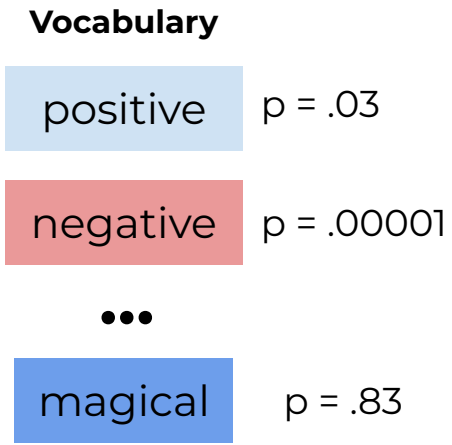
*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*



Large Language Models (LLMs)

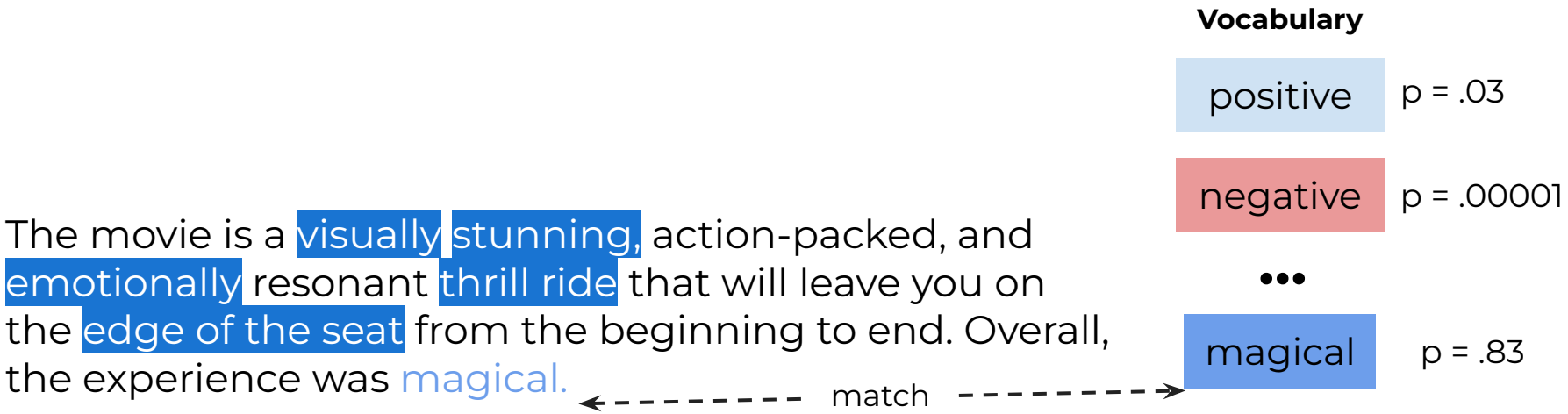
*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*

The movie is a visually stunning, action-packed, and emotionally resonant thrill ride that will leave you on the edge of the seat from the beginning to end. Overall, the experience was magical.



Large Language Models (LLMs)

*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*



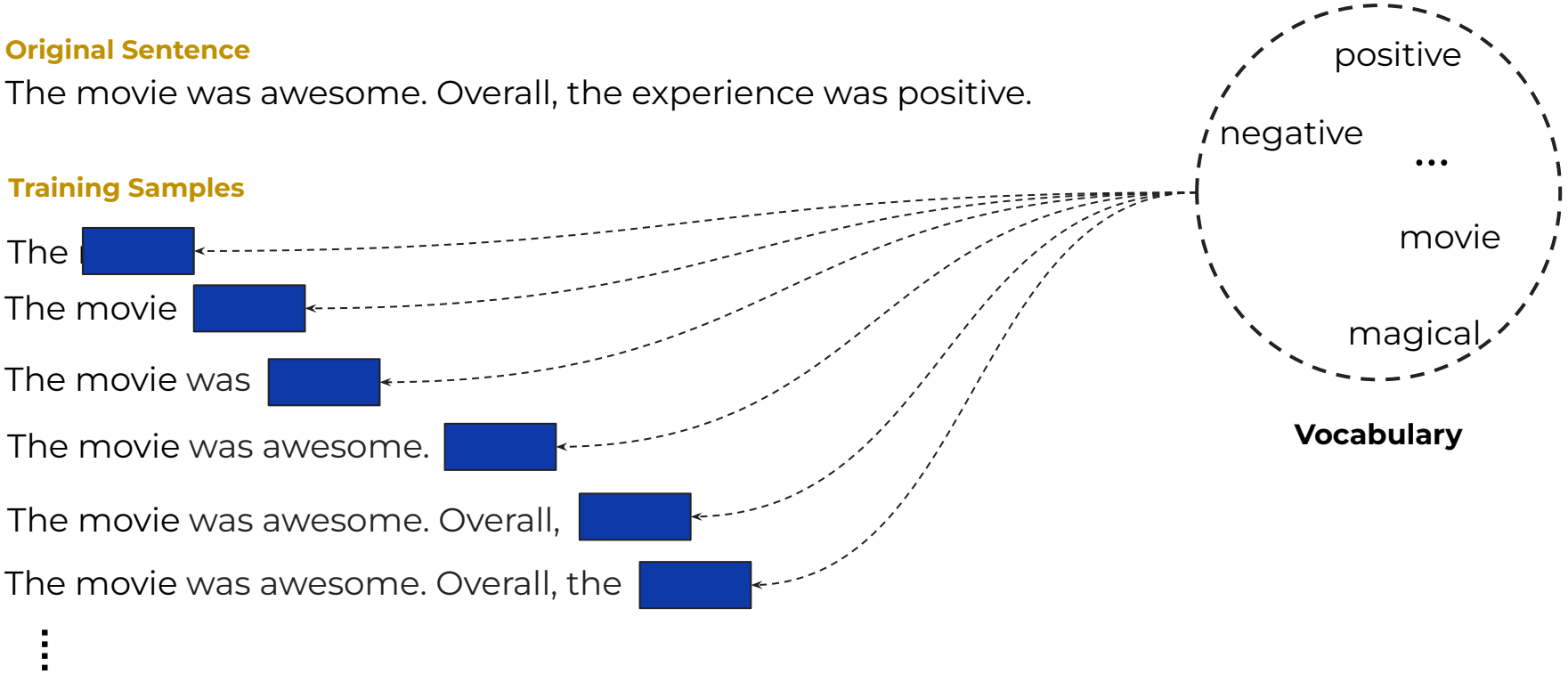
Large Language Models (LLMs)

*LLMs are **train**ed using language modeling, that is, predicting the next word in a sequence.*

Original Sentence

The movie was awesome. Overall, the experience was positive.

Training Samples



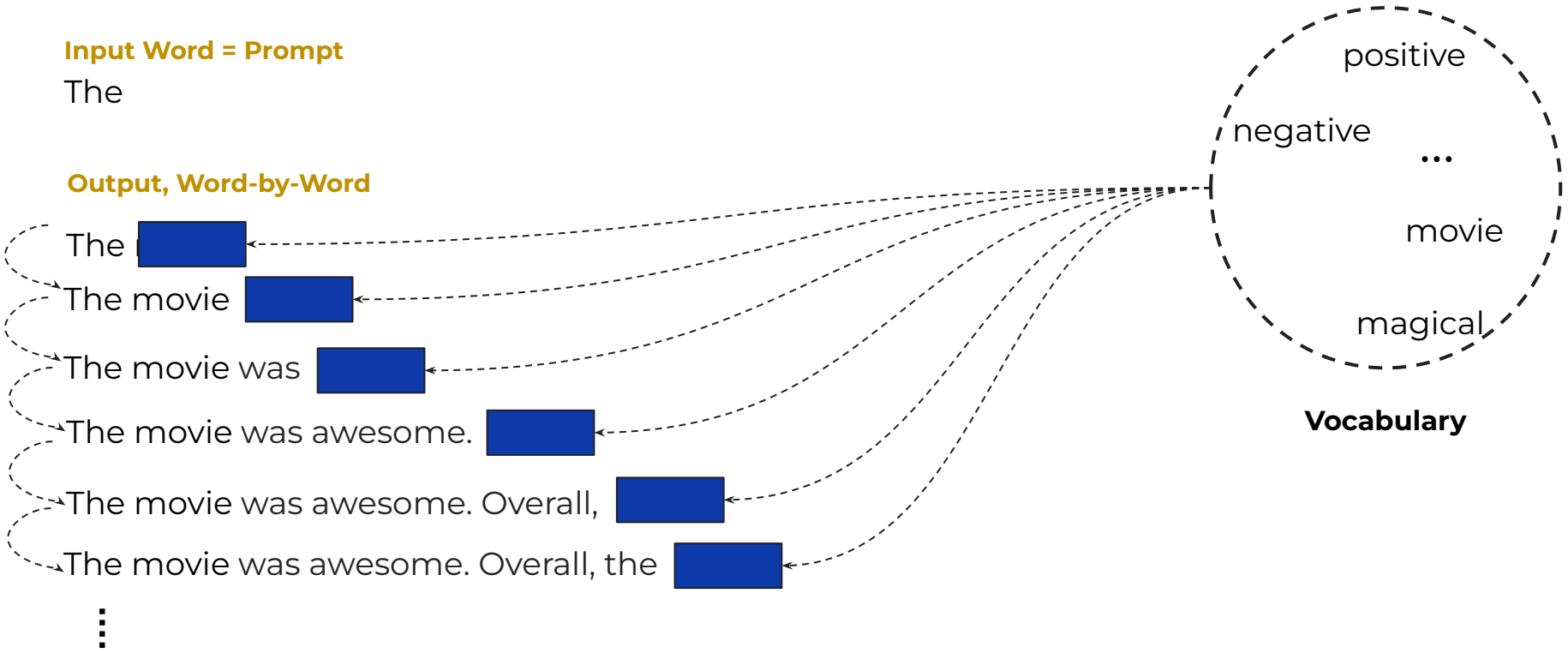
Large Language Models (LLMs)

*During **inference**, the LLM predicts the next word in the input sequence.*

Input Word = Prompt

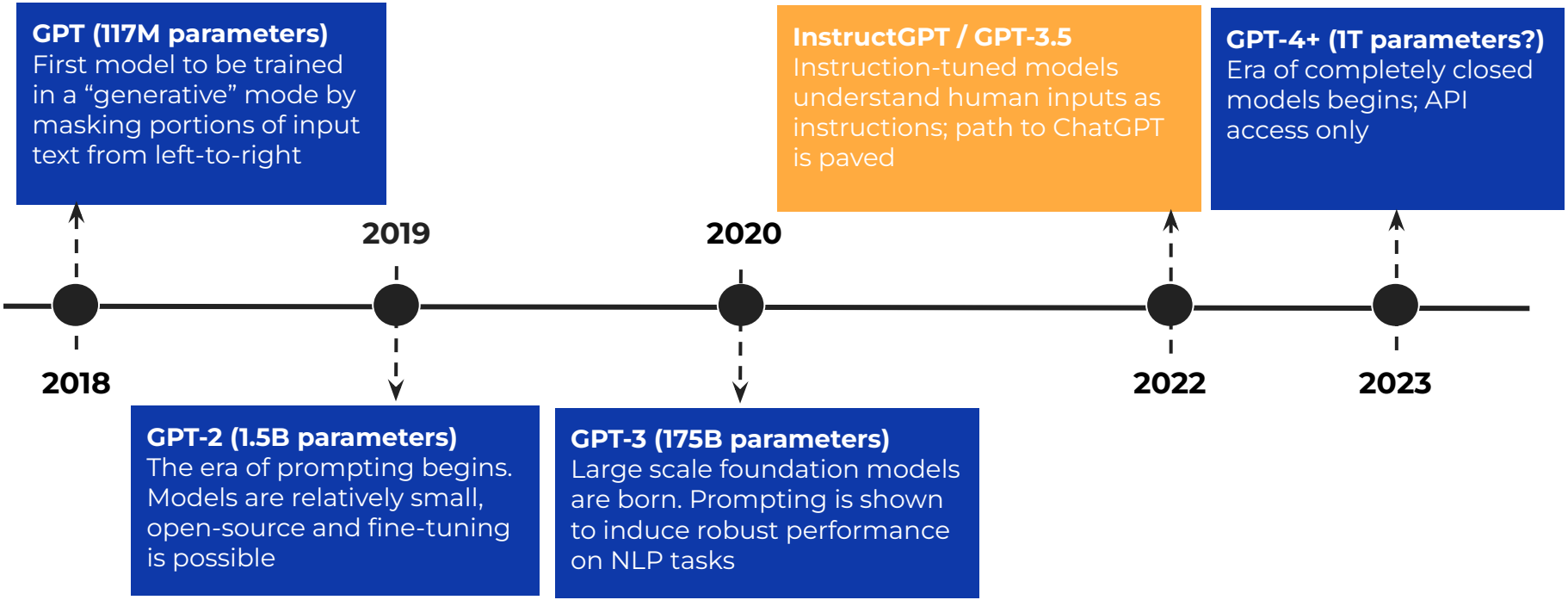
The

Output, Word-by-Word



Large Language Models (LLMs)

*Over the last 2 years, LLMs like ChatGPT have evolved to become **Conversational AI assistants***



The OpenAI GPT Assistant Training Recipe
High-Quality Assistant-style Chat Models

- **OpenAI's GPT-3.5** (behind the original ChatGPT of Nov 2022) and **GPT-4**, based on OpenAI's Generative Pre-trained Transformer (GPT) series, **first pioneered the multi-stage training procedure** for creating the **highest quality Instruction-tuned Chat Assistants**

  - honest, helpful and harmless by being aligned with human values.

OpenAI's GPT Assistant Training Recipe

Stage	**Pretraining**	**Supervised Fine-tuning**	**Reward Modeling**	**Reinforcement Learning**
Dataset	**Raw Internet** Trillions of words / tokens Low quality, Large quantity	**Ideal Assistant Responses** 10-100K words / tokens High quality, Low quantity	**Comparisons** 100K-1M comparisons High quality, Low quantity	**Prompts** 10K-100K prompts High quality, Low quantity
Training Objective	**Language Modeling** Predict the next word	**Language Modeling** Predict the next word	**Binary Classification** Predict rewards consistent with preferences	**Reinforcement Learning** Generate words that maximize reward
Model	**Base Model**	**SFT Model**	**Reward Model**	**RLHF Model**

Summary

So in order to summarize:

- We got a glimpse into what the term Large Language Model (LLM) means, and we understood how to differentiate between Base LLMs and Chat LLMs
- We understood the training regimen and the next-word prediction objective behind Autoregressive LLMs, which governs how LLMs are trained to generate text one word at a time
- We also made the link from training to inference, and saw how the LLM inference process is merely the forward propagation step of the training process
- We also understood how LLMs (with this same Autoregressive Core Mechanism) have evolved from mere Text Completion Systems to Instruction-tuned Conversational AI Assistants
- Finally, we went through the modern multi-stage GPT Training Recipe pioneered by OpenAI's GPT series of LLMs, and understood each step of that multi-stage process in greater detail