ChatGPT and how it works – my collection of notes:

# Introduction

What is ChatGPT?

ChatGPT is a language model that takes a user prompt and generates a textual response. The responses generated by ChatGPT are the most realistic that are seen by any chatbot existing in the world until today. Understanding ChatGPT requires the understanding of the following concepts:

* The GPT – that is – the Generative Pre-trained transformers which are the language models built on top of the Transformer Neural Network
* Reinforcement Learning

Diagram

Description automatically generated

**Figure: The Transformer Architecture**

Let us understand the above ingredients of ChatGPT in some detail:

Firstly,

Language Models – What are Language Models?

Language models are models which have an inherent understanding of language in a mathematical sense. Language models understand the probability distribution of a sequence of words. That is: given a context of words that have preceded it, language models will know what word (or, precisely the token) to generate next. One can get language models to handle specific tasks like:

* Question and Answering
* Text Summarization
* Machine Translation
* More.

Some of the language models include the following:

1. Bidirectional Encoding Representation of Transformers [BERT]
2. Generative Pretrained Transformers [GPT]
3. Recurrent Neural Networks [RNN]
4. Long Short-Term Memory [LSTM] networks

The Language models are pre-trained on general language data and then fine-tuned depending upon the task we want to solve. ChatGPT is a Generative Pretrained Transformer (GPT) model that is fine tuned to respond to a user’s request and then fine tuned further by reinforcement learning. I will discuss about reinforcement learning in a subsequent post.

# How is ChatGPT trained?

Let us go a bit deeper into ChatGPT. ChatGPT has been trained using Reinforcement Learning from Human Feedback (RLHF). The training of ChatGPT can be summarized into 3 basic steps:

**Step 1:**

An initial model of ChatGPT was trained using supervised fine tuning (SFT) – in which there was already a pre-trained language model – the pre-trained language model in case of ChatGPT was Generative Pre-Trained Transformer (GPT).

Fine tuning of GPT has been carried out using human AI trainers that provided the questions and answers. That is: the AI trainers – the labellers write a prompt and also write the response to the answer of the prompts – this is the supervised fine tuning of the GPT model.

**Step 2:**

In the next step of ChatGPT training, a comparison data was collected which consisted of 2 or more “model” responses ranked by quality. In this case, the responses were that of the fine tuned GPT model but were ranked by human AI trainers.

This prompt / response / ranking is used to train another GPT model which is called the “Rewards Model”. The input to the “Rewards Model” is the prompt and one of the responses by the supervised fine tined model of Step 1 and the output is going to be a reward which is the quantification of how good the response was.

**Step 3:**

In the step 3 of the GPT training, an unseen prompt is taken and passed through a copy of the supervised fine-tuned model and then through the rewards model of step 2 to get a rank.

Whatever rank is obtained is used to further fie tune the model to create a better response. This process helps the GPT model to create more factual responses, more coherent responses with hat of the question and eliminating any biased or adverse responses.

Graphical user interface, application

Description automatically generated

**Figure: The 3 steps in ChatGPT Training**