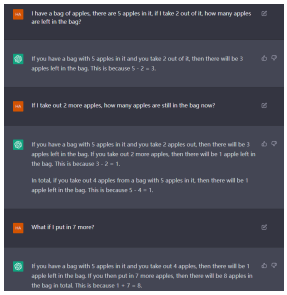
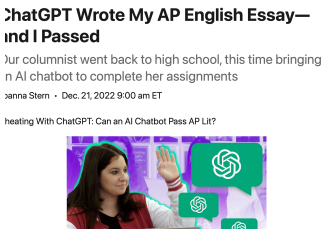
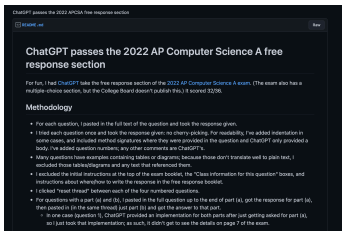
ChatGPT

VCHS AI Club







Many school districts are still scrambling to figure out how to set policies on if and how it can be used.

The New York City education department said that it's restricting access on school networks and devices because it's worried about negative impacts on student learning, as well as "concerns regarding the safety and accuracy of content."

> *"To determine if something was written by a human or an AI, you can look for the absence of personal experiences or emotions, check for inconsistency in writing style, and watch for the use of filler words or repetitive phrases. These may be signs that the text was generated by an AI."*

That's what ChatGPT told an AP reporter when asked how to tell the difference.

OpenAI said in a human-written statement this week that it plans to work with educators as it learns from how people are experimenting with ChatGPT in the real world.

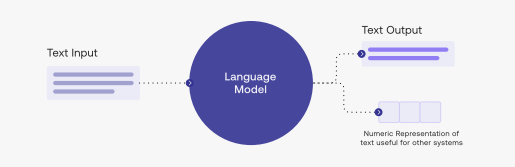
What is ChatGPT?

ChatGPT is the latest language model from OpenAI and represents a significant improvement over its predecessor GPT-3.

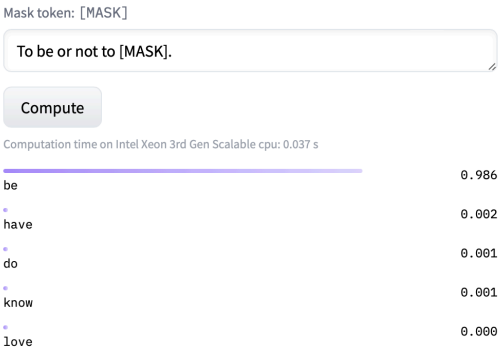
Similarly to many Large Language Models, ChatGPT is capable of generating text in a wide range of styles and for different purposes, but with remarkably greater precision, detail, and coherence.

It represents the next generation in OpenAI's line of Large Language Models, and it is designed with a strong focus on interactive conversations.

What is a Large Language Model?



Large Language Models (LLMs) are artificial intelligence tools that can read, summarize and translate texts and predict future words in a sentence letting them generate sentences similar to how humans talk and write.



Masked Language Modeling

They are neural networks that have been trained on hundreds of gigabytes of text gathered from the web. During training, the network is fed with text excerpts that have been partially masked. The neural network tries to guess the missing parts and compares its predictions with the actual text.

For example, if the model is given the input sentence

"The cat sat on the"

it might predict the next word as "mat", "chair", or "floor" because of the high-probability of occurrence of these words given the previous context; the language model is in fact able to estimate the likelihood of each possible word (in its vocabulary) given the previous sequence.

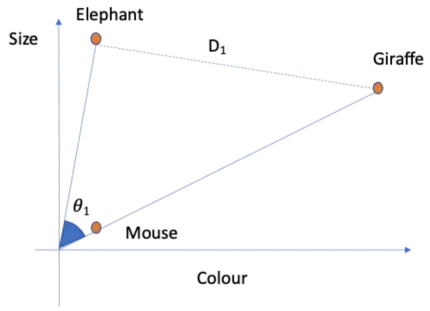
The masked language modeling approach is a variant of next token prediction, in which some of the words in the input sentence are replaced with a special token, such as [MASK]. The model is then asked to predict the correct word that should be inserted in place of the mask.

For example, if the model is given the sentence

"The [MASK] sat on the"

as input, it might predict the next word as "cat", "dog", or "rabbit".

One advantage of these objective functions is that it allows the model to learn the statistical structure of language, such as common word sequences and patterns of word usage. This generally helps the model generate more natural and fluent text, and it is an essential step in the pre-training phase of every language model.



Comparing some vectors in a 2D space using Euclidian and Cosine distance

While traditional Natural Language Processing (NLP) algorithms typically only look at the immediate context of words, LLMs consider large swaths of text in order to better understand the context.

For example, if a sentence reads ‘She was walking with an umbrella,' an LLM can better understand the meaning and context of ‘umbrella’ by looking at sentences surrounding it. This allows LLMs to more accurately interpret the data they are presented with.

By using large pre-trained models, such as different versions of GPT-3 or BERT, Large Language Models can provide high levels of accuracy in areas from sentiment analysis to natural language generation. They can even be used for tasks as complex as summarization and question answering.

What's the problem?

However, these objective functions can also lead to problems, essentially because the model is not capable of distinguishing between an important error and an unimportant one.

To make a very simple example, if the model is given the input sentence:

"The Roman Empire [MASK] with the reign of Augustus."

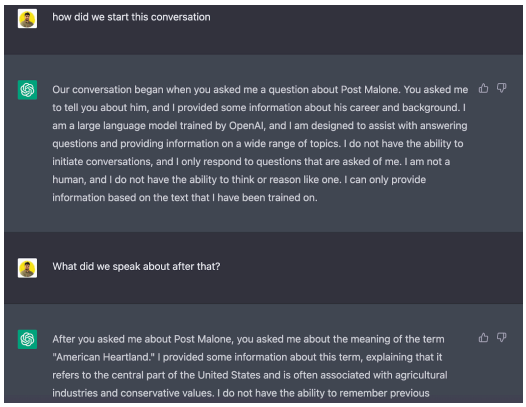
it

might predict "began" or “ended”, as both words score high likelihood of occurrence (indeed, both sentences are historically correct), even though the second choice implies a very different meaning.

More generally, these training strategies can lead to a misalignment of the language model for some more complex tasks, because a model which is only trained to predict the next word (or a masked word) in a text sequence, may not necessarily be learning some higher-level representations of its meaning. As a result, the model struggles to generalize to tasks or contexts that require a deeper understanding of language.

What makes ChatGPT unique?

The creators have used a combination of both Supervised Learning and Reinforcement Learning to fine-tune ChatGPT, but it is the Reinforcement Learning component specifically that makes ChatGPT unique. The creators use a particular tech-



nique called Reinforcement Learning from Human Feedback (RLHF), which uses human feedback in the training loop to minimize harmful, untruthful, and/or biased outputs.

Researchers and developers are working on various approaches to address the alignment problem in Large Language Models. ChatGPT is based on the original GPT-3 model, but has been further trained by using human feedback to guide the learning process with the specific goal of mitigating the model's misalignment issues. The specific technique used, called Reinforcement Learning from Human Feedback, is based on previous academic research. ChatGPT represents the first case of use of this technique for a model put into production.

Reinforcement Learning from Human Feedback

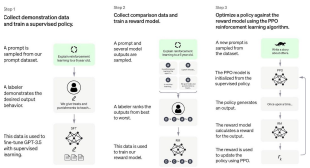
THE METHOD OVERALL CONSISTS OF THREE DISTINCT STEPS:

1. Supervised fine-tuning step: a pre-trained language model is fine-tuned on a relatively small amount of demonstration data curated by labelers, to learn a supervised policy (the SFT model) that generates outputs from a selected list of prompts. This represents the baseline model.

2. "Mimic human preferences" step: labelers are asked to vote on a relatively large number of the SFT model outputs, this way creating a new dataset consisting of comparison data. A new model is trained on this dataset. This is referred to as the reward model (RM).

3. Proximal Policy Optimization (PPO) step: the reward model is used to further fine-tune and improve the SFT model. The outcome of this step is the so-called policy model.

Step 1 takes place only once, while steps 2 and 3 can be iterated continuously: more comparison data is collected on the current best policy model, which is used to train a new reward model and then a new policy.



Performance Evaluation

Because the model is trained on human labelers input, the core part of the evaluation is also based on human input, i.e., it takes place by having labelers rate the quality of the model outputs. To avoid overfitting to the judgment of the labelers involved in the training phase, the test set uses prompts from held-out OpenAI customers which are not represented in the training data.

The model is evaluated on three high-level criteria:

– Helpfulness: judging the model's ability to follow user instructions, as well as infer instructions.

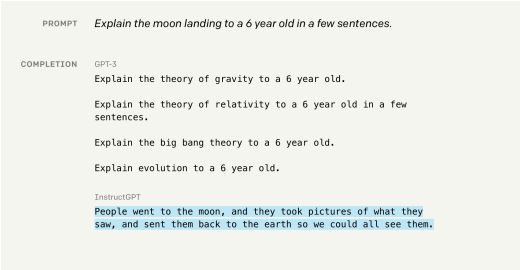
– Truthfulness: judging the model's tendency for hallucinations (making up facts) on closed-domain tasks. The model is evaluated on the TruthfulQA dataset.

– Harmlessness: the labelers evaluate whether the model's output is appropriate, denigrates a protected class, or contains derogatory content.

The model is also benchmarked on the RealToxicityPrompts and CrowS-Pairs datasets. The model is also evaluated for zero-shot performance on traditional NLP tasks like question answering, reading comprehension, and summarization, on some of which the developers observed performance regressions compared to GPT-3. This is an example of an "alignment tax" where the RLHF-based alignment procedure comes at the cost of lower performance on certain tasks.

The performance regressions on these datasets can be greatly reduced with a trick called pre-train mix: during training of the PPO model via gradient descent, the gradient updates are computed by mixing the gradients of the SFT model and the PPO model.

InstructGPT



The OpenAI API is powered by GPT-3 language models which can be coaxed to perform natural language tasks using carefully engineered text prompts. But these models can also generate outputs that are untruthful, toxic, or reflect harmful sentiments. This is in part because GPT-3 is trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants. In other words, these models aren't aligned with their users.

To make our models safer, more helpful, and more aligned, we use an existing technique called reinforcement learning from human feedback (RLHF). On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.

The resulting InstructGPT models are much better at following instructions than GPT-3. They also make up facts less often, and show small decreases in toxic output generation. Our labelers prefer outputs from our 1.3B InstructGPT model over outputs from a 175B GPT-3 model, despite having more than 100x fewer parameters. At the same time, we show that we don't have to compromise on GPT-3's capabilities, as measured by our model's performance on academic NLP evaluations.