AI for Education

Flint

How does AI know stuff?

Generative AI models (like the ones that power ChatGPT) get their knowledge from **training** and then from any **context** that you provide them.

Pretraining

Generative AI (GenAI) models start their learning from hundreds of terabytes of data from the internet. For example, ChatGPT 3.5 was trained on 175 billion parameters. This initial step is called "**pretraining**."

- Companies like OpenAI spent months and millions of dollars on this stage. For instance, the GenAI model behind ChatGPT cost OpenAI over $100 million just in this phase!
- During pretraining, models learn from websites such as Wikipedia, books, and online articles.
- As a result, they can communicate in various languages, analyze data, and even write computer code.

Further tuning

After pretraining, models undergo "**instruction tuning**" and sometimes "**Reinforcement Learning from Human Feedback (RHLF)**." These steps make the AI's responses more useful to users.

- Although these steps don't add new facts, they teach the model how to better answer questions and prompts. Most modern models have at least one of these additional steps.
- Both of these tuning phases use data labeled by humans to refine the AI's behavior.

Limitations and Safeguarding

GenAI models have knowledge limits. For instance, ChatGPT 3.5 is unaware of events after September 2021.

- Since GenAI models learn from internet data, they might echo its biases.
- There's a risk of these models creating or "hallucinating" information.
- However, the following **safeguarding techniques** are integrated during training, ensuring the AI chooses its responses more cautiously:

Safeguarding Techniques

Filtered Training Data

Before training, the data used is filtered to remove harmful or inappropriate content. This process helps in reducing the chances of the model generating unsafe outputs.

Human Feedback

During the tuning phases, human reviewers assess and rate potential model outputs for a range of example inputs. Feedback from these reviewers is crucial in refining the model's behavior.

Regular Updates

The model is regularly updated based on user feedback and observed behavior to ensure it aligns with safety and ethical guidelines.

External Audits

Some organizations conduct third-party audits on their models and training processes to ensure that best practices for safety are being followed.

Context Windows

When you talk to ChatGPT, it's like it has a "cheat sheet" in front of it. This note contains the recent things you've said and its own replies. ChatGPT looks at this note to reply to you. However, it can only see a limited amount of words at once, so if the conversation gets too long, it might forget the earlier parts. It doesn't "remember" like humans do; it just refers back to its cheat sheet.

- The term "**context**" refers to this note – it's the combined list of recent things both you and the AI have said.
- Each GenAI model has a different amount of information it can remember in a context window:
  - For example, ChatGPT can remember the last ~3,000 words in a conversation.
  - New AI models like Claude can remember up to 75,000 words, or about the length of “Harry Potter and the Sorcerer’s Stone” at a time. That said, AI models with larger context windows tend to have lower quality responses.

Sources

“GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery”

“Introducing ChatGPT”

“Improving Language Understanding by Generative Pre-Training”

__AI for Education__ helps educators and schools responsibly adopt AI technology, empowering teachers and ultimately improving student outcomes while preparing them for the future. Learn more at __aiforeducation.io__

__Flint__ is an AI platform built for schools that doesn't use student chat data to train AI models. It also enables admins to control access, view usage trends, and request student chat history. Learn more at __flintk12.com__