PDFs store highlighted text data differently compared to regular text. When text is highlighted in a PDF, the highlighting is typically represented as an annotation rather than altering the actual text content. These annotations contain metadata about the highlight, such as its position and shape, but they do not inherently carry the highlighted text itself. Instead, the process of extracting highlighted text involves analyzing the positions of the highlights and cross-referencing them with the text content of the PDF to determine which text was highlighted.

From the provided sources, here's a summary of how highlighted text data is stored and how it can be extracted:

Sticky Notes vs. Highlights: Sticky notes added by Adobe Reader are easier to parse because they append both content and position information to the PDF. However, for highlights, there is only rectangle information available, indicating the area of the text that was highlighted. To extract the text, one needs to locate the text within the highlighted area based on its position 2.

ChatGPT is a chatbot and virtual assistant developed by OpenAI and launched on November 30, 2022. Based on large language models (LLMs), it enables users to refine and steer a conversation towards a desired length, format, style, level of detail, and language. Successive user prompts and replies are considered at each conversation stage as context.[2]

ChatGPT is credited with starting the AI boom, which has led to ongoing rapid investment in and public attention to the field of artificial intelligence.[3] By January 2023, it had become what was then the fastest-growing consumer software application in history, gaining over 100 million users and contributing to the growth of OpenAI's current valuation of $86 billion.[4][5] ChatGPT's release spurred the release of competing products, including Gemini, Claude, Llama, Ernie and Grok.[6] Microsoft launched Copilot, now based on OpenAI's GPT-4o. Some observers raised concern about the potential of ChatGPT and similar programs to displace or atrophy human intelligence, enable plagiarism, or fuel misinformation.[7][8]

ChatGPT is built on OpenAI's proprietary series of generative pre-trained transformer (GPT) models and is fine-tuned for conversational applications using a combination of supervised learning and reinforcement learning from human feedback.[7] ChatGPT was released as a freely available research preview, but due to its popularity, OpenAI now operates the service on a freemium model. Users on its free tier can access GPT-4o and GPT-3.5. The ChatGPT subscriptions "Plus", "Team" and "Enterprise" provide additional features such as DALL-E 3 image generation and increased GPT-4o usage limit.[9]

Training

ChatGPT is based on particular GPT foundation models, namely GPT-3.5 and GPT-4, that were fine-tuned to target conversational usage.[10] The fine-tuning process leveraged supervised learning and reinforcement learning from human feedback (RLHF).[11][12] Both approaches employed human trainers to improve model performance. In the case of supervised learning, the trainers played both sides: the user and the AI assistant. In the reinforcement learning stage, human trainers first ranked responses that the model had created in a previous conversation.[13] These rankings were used to create "reward models" that were used to fine-tune the model further by using several iterations of proximal policy optimization.[11][14]

Time magazine revealed that, to build a safety system against harmful content (e.g., sexual abuse, violence, racism, sexism), OpenAI used outsourced Kenyan workers earning less than $2 per hour to label harmful content. These labels were used to train a model to detect such content in the future. The outsourced laborers were exposed to "toxic" and traumatic content; one worker described the assignment as "torture". OpenAI's outsourcing partner was Sama, a training-data company based in San Francisco, California.[15][16]

ChatGPT initially used a Microsoft Azure supercomputing infrastructure, powered by Nvidia GPUs, that Microsoft built specifically for OpenAI and that reportedly cost "hundreds of millions of dollars". Following ChatGPT's success, Microsoft dramatically upgraded the OpenAI infrastructure in 2023.[17] Scientists at the University of California, Riverside, estimate that a series of prompts to ChatGPT needs approximately 500 milliliters (18 imp fl oz; 17 U.S. fl oz) of water for Microsoft servers cooling.[18] TrendForce market intelligence estimated that 30,000 Nvidia GPUs (each costing approximately $10,000–15,000) were used to power ChatGPT in 2023.[19][20]

OpenAI collects data from ChatGPT users to train and fine-tune the service further. Users can upvote or downvote responses they receive from ChatGPT and fill in a text field with additional feedback.[21][22]

ChatGPT's training data includes software manual pages, information about internet phenomena such as bulletin board systems, multiple programming languages, and the text of Wikipedia.[23][24][7]

Features and limitations

Features

Although a chatbot's core function is to mimic a human conversationalist, ChatGPT is versatile. It can write and debug computer programs;[25] compose music, teleplays, fairy tales, and student essays; answer test questions (sometimes, depending on the test, at a level above the average human test-taker);[26] generate business ideas;[27] write poetry and song lyrics;[28] translate and summarize text;[29] emulate a Linux system; simulate entire chat rooms; play games like tic-tac-toe; or simulate an ATM.[23]

Compared to its predecessor, InstructGPT, ChatGPT attempts to reduce harmful and deceitful responses.[30] In one example, whereas InstructGPT accepts the premise of the prompt "Tell me about when Christopher Columbus came to the U.S. in 2015" as truthful, ChatGPT acknowledges the counterfactual nature of the question and frames its answer as a hypothetical consideration of what might happen if Columbus came to the U.S. in 2015, using information about the voyages of Christopher Columbus and facts about the modern world—including modern perceptions of Columbus's actions.[11]

ChatGPT remembers a limited number of previous prompts in the same conversation. Journalists have speculated that this will allow ChatGPT to be used as a personalized therapist.[31] To prevent offensive outputs from being presented to and produced by ChatGPT, queries are filtered through the OpenAI "Moderation endpoint" API (a separate GPT-based AI).[32][33][11][31]

In March 2023, OpenAI added support for plugins for ChatGPT.[34] This includes both plugins made by OpenAI, such as web browsing and code interpretation, and external plugins from developers such as Expedia, OpenTable, Zapier, Shopify, Slack, and Wolfram.