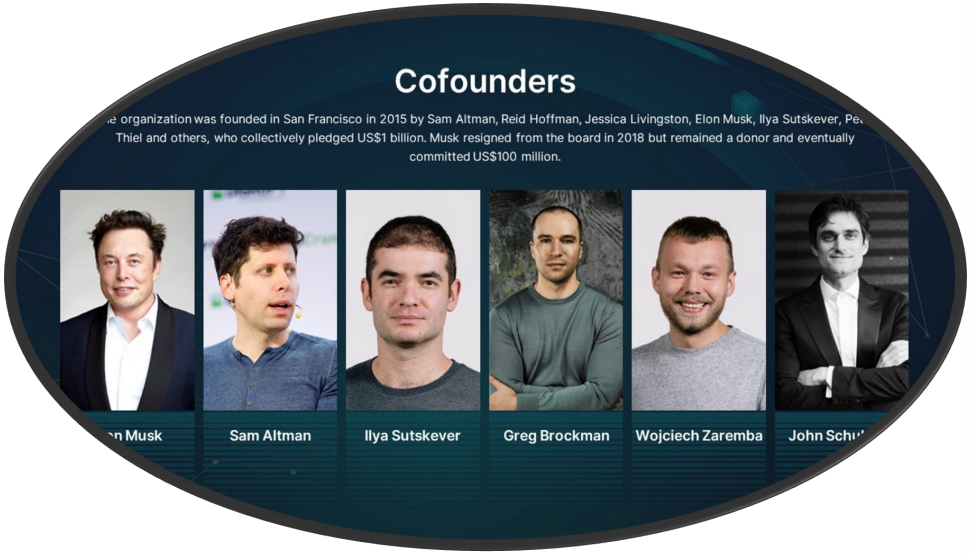
OpenAI / ChatGPT / APIs



What is OpenAI ?

Research Organization: OpenAI is an AI research lab that focuses on developing and promoting friendly AI in a way that benefits humanity as a whole.

Founded in 2015: OpenAI was established in December 2015 by Elon Musk, Sam Altman, Greg Brockman, Ilya Sutskever, Wojciech Zaremba, and others

Advanced AI Models: Developing some of the most advanced AI models, including the Generative Pretrained Transformer (GPT) series, with the latest iterations being GPT3 and GPT4.

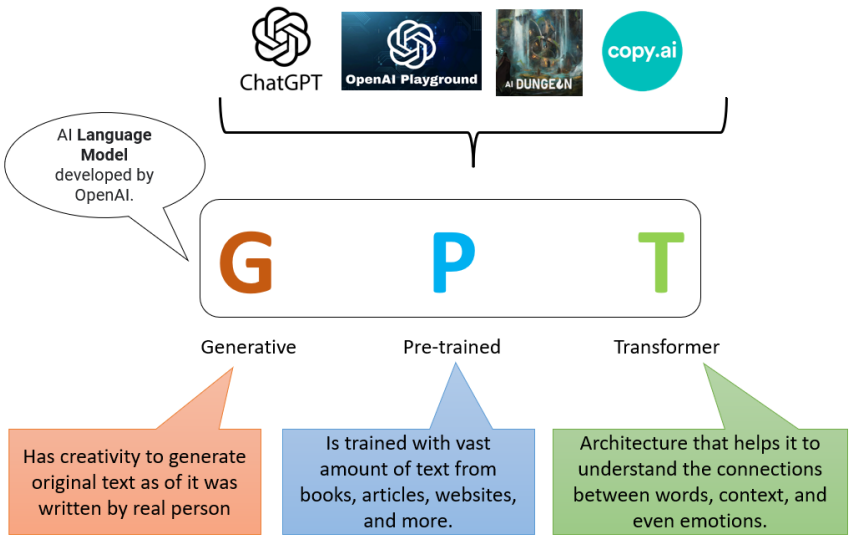
Ethics and Safety: A significant focus for OpenAI is ensuring the ethical use of AI and addressing potential safety risks associated with powerful AI systems.

Open Collaboration: Initially, OpenAI started with a commitment to openaccess research, sharing its findings and technologies. However, it has since adopted a more controlled release approach to mitigate potential risks.

Commercial Products: OpenAI has developed several commercial products, such as the API that provides access to models like GPT3 for a variety of textbased tasks, including conversation, summarization, translation, and more.

Partnerships and Licensing: OpenAI has entered into partnerships and licensing agreements, notably with Microsoft, which provides significant Azure cloud computing resources necessary for training and deploying AI models.

What is GPT ?



1. **GPT**, which stands for "Generative Pretrained Transformer," is an advanced type of AI model developed by OpenAI for natural language processing tasks. Here are some key points about GPT:
2. **Type of AI Model:** GPT is a type of large language model (LLM) that uses deep learning techniques, specifically a transformer architecture, for understanding and generating human language.
3. **Training Method:** It is pretrained on a vast corpus of text data sourced from the internet, including websites, books, and other written material, allowing it to learn a wide range of language patterns and styles.
4. **Generative Capabilities:** GPT is designed to generate coherent and contextually relevant text based on the input it receives, making it useful for applications like content creation, conversation, and text completion.
5. **Versions and Evolution:** There have been several versions of GPT, with each new version (like GPT2, GPT3, GPT4, etc.) generally being larger and more capable than the last in terms of the amount of data it can process and the complexity of the tasks it can perform.
6. **Applications:** GPT is used in a variety of applications, including chatbots, writing assistants, language translation, and even in creative fields for generating art, music, and poetry.
7. **Humanlike Responses:** One of the notable features of GPT is its ability to produce responses that can closely mimic human writing styles, making its applications more natural and userfriendly.
8. **Ethical Considerations:** The deployment of GPT models raises ethical questions around topics like misinformation, privacy, and the potential impact on jobs in fields like writing and customer service.
9. **Accessibility and Use:** GPTpowered tools and services are increasingly accessible to businesses and individuals, enabling a wide range of users to leverage its capabilities for various purposes.

Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users

Source: UBS / Yahoo Finance

@EconomyApp

APP ECONOMY **INSIGHTS**

Models

Overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4 Turbo and GPT-4	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

- GPT 3, 4, 5
- DALL E images
- Embeddings

What is ChatGPT

Imp.

GPT-4 Variants by Approximate Size

Model	Params (Approx.)	Notes
GPT-4 Mini / GPT-4o 12B	~12B	Smallest, faster, cheaper, optimized ("Omni")
GPT-4o (default)	12B (similar or slightly larger)	Omni variant, optimized for speed and versatility
GPT-4 Turbo	Likely larger than GPT-4o but smaller than GPT-4	Optimized for faster inference, cost-efficient
GPT-4 (default)	100B+ (est.)	Full-sized, high-quality reasoning, slower

AI and Machine Learning: ChatGPT is built on the GPT (Generative Pretrained Transformer) architecture, which is a type of artificial intelligence model designed to generate text.

Language Understanding: It has been trained on a diverse range of internet text, enabling it to understand context, answer questions, write essays, and even create poetry or code.

Conversational Interface: ChatGPT is designed to engage in conversations, providing responses that can simulate a humanlike interaction.

Learning Capability: While it can't learn or remember information from individual user interactions in realtime, its design allows for continual updates and training by OpenAI to improve its performance and capabilities.

Versatile Applications: It can be used in various applications such as customer service bots, tutoring systems, content creation aids

Safety and Ethics: OpenAI has implemented safeguards to prevent ChatGPT from generating inappropriate or harmful content, though it's not foolproof.

Customization and Integration: Developers can integrate ChatGPT into their own applications through OpenAI's API, allowing for a wide range of custom uses and functionalities tailored to specific needs.

Continual Development: ChatGPT is part of an ongoing research and development effort, with improvements and new versions being released as AI technology advances.

GPT MODELS COMPARISON CHART

Model	Size	Memory capacity	Accuracy	Input formats	Price
GPT-3	175B	1,500 words	<60%	Text, speech	$$
GPT-3.5	20B	8,000 words	<60%	Text, speech	$
GPT-4	>1T (?)	25,000-64,000 words	>80%	Text, speech, image	$$$

greenice

CHAT GPT-3 VS. CHAT GPT-4

CHAT GPT-3	CHAT GPT-4
Only takes text prompts	Take text & image prompts
Creative...sort of	More creative
Hallucinates a lot of facts & opinions	Still Hallucinates, but not as much :)
Barely passed the bar exam	Aced the bar exam
Takes a lot of steering & prompts for developers	More steerable in conversations & developer prompts

"GPT-4 IS MORE RELIABLE, CREATIVE AND ABLE TO HANDLE MUCH MORE NUANCED INSTRUCTIONS THAN GPT-3.5" - OPEN AI

GPT3 Vs GPT4

GPT3 and GPT4 are both iterations of OpenAI's Generative Pretrained Transformer (GPT) series, but there are key differences between them:

1.Model Size and Complexity:

- **GPT3:** Has 175 billion parameters, making it one of the largest language models at its time of release.
- **GPT4:** Is even larger than GPT3, with more parameters (the exact number has not been publicly disclosed), which enhances its understanding and generation capabilities.

2.Performance and Accuracy:

- **GPT3:** Sometimes struggles with complex reasoning tasks and can generate less accurate or relevant responses in certain contexts.
- **GPT4:** Shows improved performance in understanding context and nuance, leading to more accurate and contextually relevant outputs.

3.Training Data and Knowledge:

- **GPT3:** Trained on a vast corpus of text data available up until its training cutoff in 2020.
- **GPT4:** Benefits from an even larger and more diverse dataset, including more recent information, leading to a broader range of knowledge and understanding.

4. Capabilities in Understanding Context:

- **GPT3:** Exhibits a good understanding of context but can lose coherence over longer conversations or more complex queries.
- **GPT4:** Demonstrates a significantly improved ability to maintain context over longer interactions

5. Multimodal Abilities:

- **GPT3:** Primarily focused on text processing and generation.
- **GPT4:** It can understand and generate not just text but also images

6. Application and Usage:

- **GPT3:** Widely used across various industries for tasks like content creation, coding, customer service, and more.
- **GPT4:** Expands on these applications with improved performance, making it more effective and versatile for complex tasks.

7. Error Rates and Reliability:

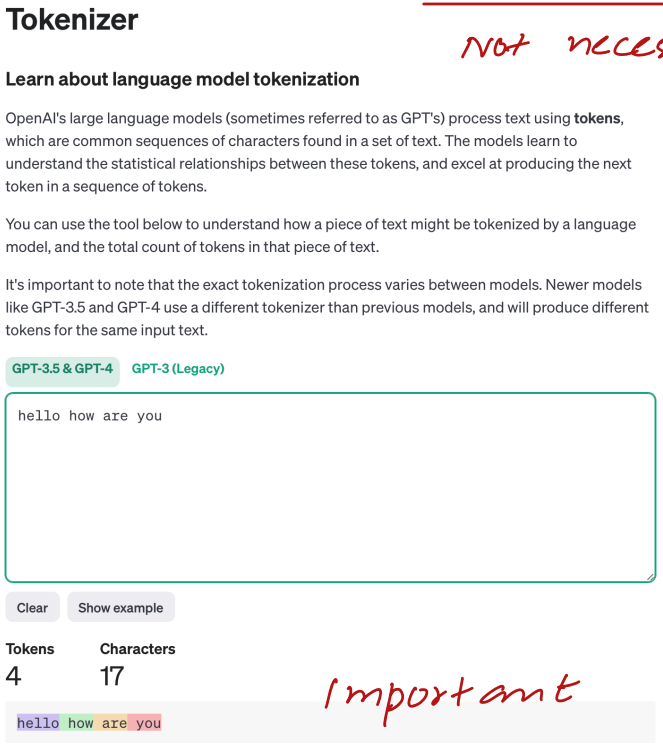
- **GPT3:** Though advanced, it has higher error rates in certain complex tasks.
- **GPT4:** Demonstrates a lower error rate and higher reliability in a wide range of tasks.

Playing

2 Tokens ; play + ing

not necessarily words

What are Tokens ?



Important

Input

Output

Token } Separate pricing

Basic Units of Text: In NLP, a token is the basic unit of text. It can be a word, a part of a word (like a prefix or suffix), or even punctuation. Tokens are the building blocks that models like ChatGPT analyze and generate.

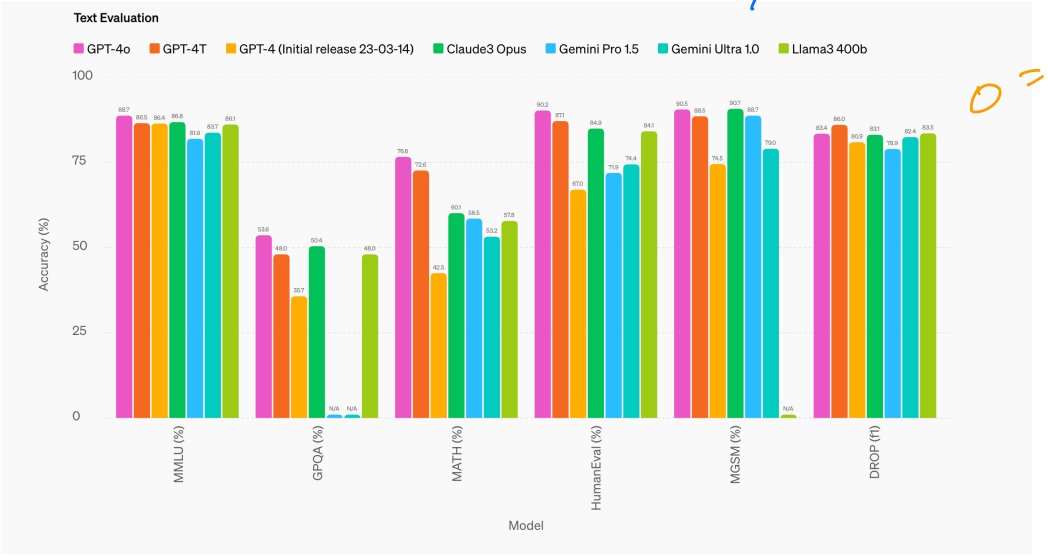
Tokenization: This is the process of breaking down text into its constituent tokens.

Sequence of Tokens: When you input a sentence, ChatGPT tokenizes it, processes the tokens to understand the context and generate a response, and then converts the output tokens back into humanreadable text.

Vocabulary Limit: Language models have a fixed vocabulary size, meaning they can only recognize a certain number of unique tokens.

Efficiency in Processing: Using tokens allows language models to efficiently process large amounts of text by breaking down complex structures into manageable pieces.

GPT 4 = 175B params



O = Omni

GPT-4o

only 12B

GPT-4O is an advanced version of the GPT-4 model developed by OpenAI. The "O" stands for "Omni," highlighting its all-encompassing capabilities.

Release Date : May 2024

Multimodal Input (Omni): GPT-4O can process and understand images in addition to text, unlike GPT-3 which only handles text.

Context Length: GPT-4O can maintain context over longer conversations compared to GPT-3.

Higher Accuracy in Response Generation: GPT-4O provides more accurate and relevant responses, reducing the chances of generating incorrect or nonsensical text.

Improved Efficiency and Speed: GPT-4O is optimized to deliver responses faster than GPT-3, making it more suitable for real-time applications.

Enhanced Adaptability: GPT-4O can better adapt to different styles and tones based on user input, providing more personalized interactions.

Human like Response Time : It can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds

Pricing : It's considered to be almost 50% cheaper as compared to GPT-4 Models

Model	Pricing
gpt-4o	US$5.00 / 1M input tokens US$15.00 / 1M output tokens

→ Faster (less accuracy)

half cost

Input, Output } Both

Model	Input	Output
gpt-4-turbo	US$10.00 / 1M tokens	US$30.00 / 1M tokens
gpt-4-turbo-2024-04-09	US$10.00 / 1M tokens	US$30.00 / 1M tokens

Small Language Models

GPT-4o Mini

SLM

Small – Compact

GPT-4o Mini is a compact, efficient version of the larger GPT-4 model designed for deployment in resource-constrained environments.

Release Date : July 2024

Compact Size
GPT-4o Mini is designed to run efficiently on hardware with limited computational resources.

Faster Inference
Provides faster response times compared to full-sized GPT models, making it ideal for time-sensitive tasks.

Energy Efficiency
Prioritizes low power consumption, unlike larger models that require significant energy to function.

Edge Deployment
Designed to function independently of cloud infrastructure, unlike traditional GPT models that rely heavily on server-based computations.

Customizable and Scalable
Offers greater flexibility in adapting to particular use cases, unlike one-size-fits-all larger GPT models.

Cost-Effective
Lower operational costs due to reduced resource requirements.

Simplified Training
Difference: Easier and faster to train compared to large-scale GPT models that need extensive datasets and powerful GPUs.

To stay in competition

Model	Pricing
gpt-4o-mini	US$0.150 / 1M input tokens US$0.600 / 1M output tokens

Model	Pricing
gpt-4o	US$5.00 / 1M input tokens US$15.00 / 1M output tokens

Model	Input	Output
gpt-4-turbo	US$10.00 / 1M tokens	US$30.00 / 1M tokens
gpt-4-turbo-2024-04-09	US$10.00 / 1M tokens	US$30.00 / 1M tokens

Open AI Models

Use Case / Modality	Model Name(s) / Family	Example Client Code	Input → Output
Chat / Text Generation	gpt-4, gpt-4o, gpt-4o-mini, gpt-3.5-turbo	client.chat.completions.create()	Text → Text
Image Generation	dall-e-2, dall-e-3, gpt-image-1	client.images.generate()	Text → Image
Speech-to-Text (ASR)	whisper-1	client.audio.transcriptions.create()	Audio → Text
Text-to-Speech (TTS)	gpt-4o-mini-tts	client.audio.speech.create()	Text → Audio
Video Generation	sora (limited access)	*(not public in API yet)*	Text → Video
Embeddings	text-embedding-ada-002, text-embedding-3-small/large	client.embeddings.create()	Text → Vector

All models are based on

- Reinforment learning + Human Response

Ph.D Level O_1 = Reasoning model (more compute)

- It uses chain of thought

Q// why is O_1 costly?

because of Input / Reasoning (extra) / output tokens

Chain of Thought (CoT)

- The model writes **step-by-step reasoning** before giving the answer.

Q: If a bat and ball cost $1.10 together, and the bat costs $1 more than the ball, how much does the ball cost?

A (CoT):

1. Let the price of the ball be x.
2. Then the bat costs x + $1.
3. Total cost = x + (x + 1) = 1.10
4. Solve: 2x + 1 = 1.10 → 2x = 0.10 → x = 0.05

   Answer: $0.05 ✅

Prompting Types

- Zero shot
- one shot
- few shot
- Chain of thoughts

Imp. OpenAI started as Non profit but now it is private entity.