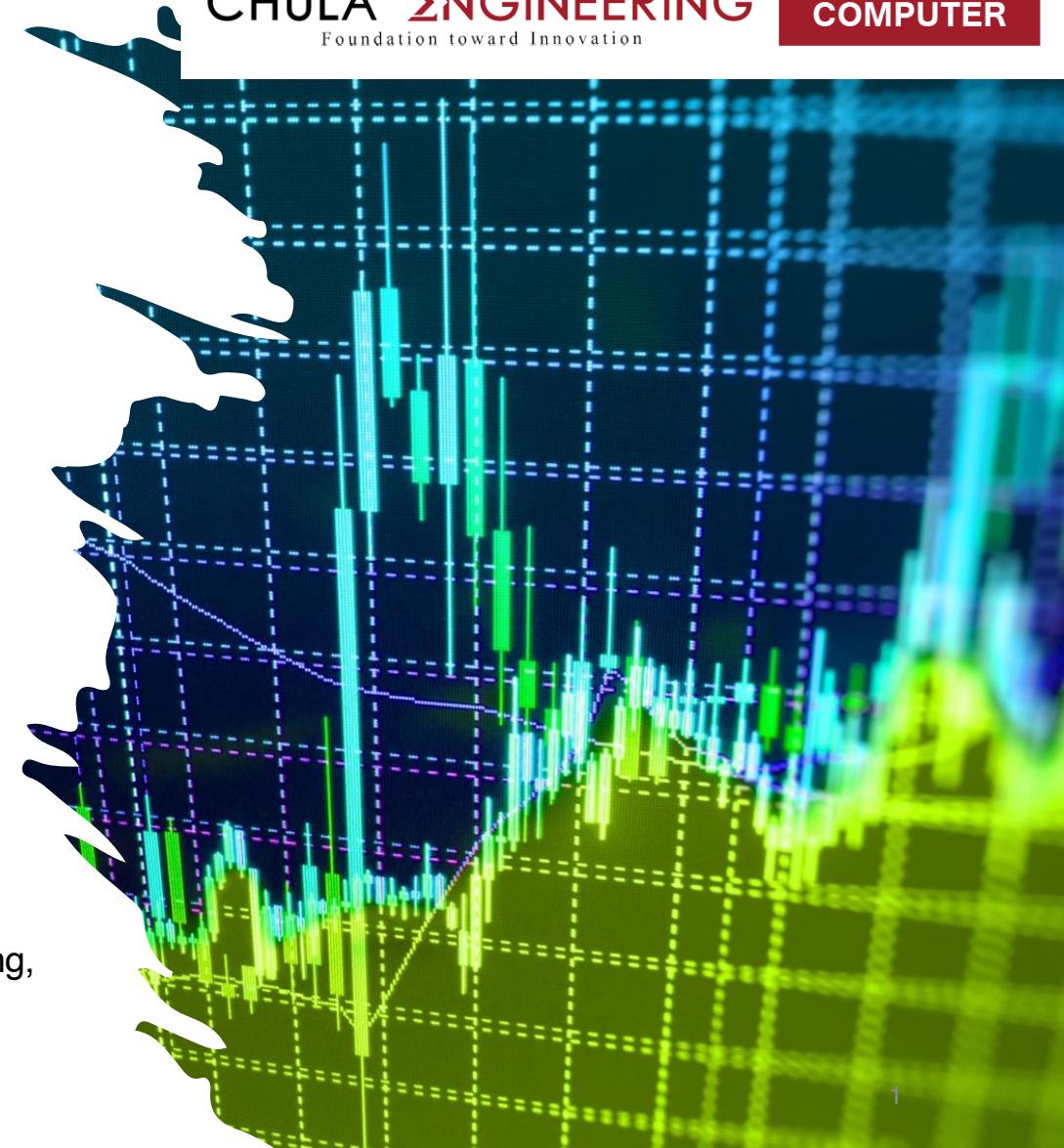


Generative AI

Assoc. Prof. Peerapon Vateekul, Ph.D.

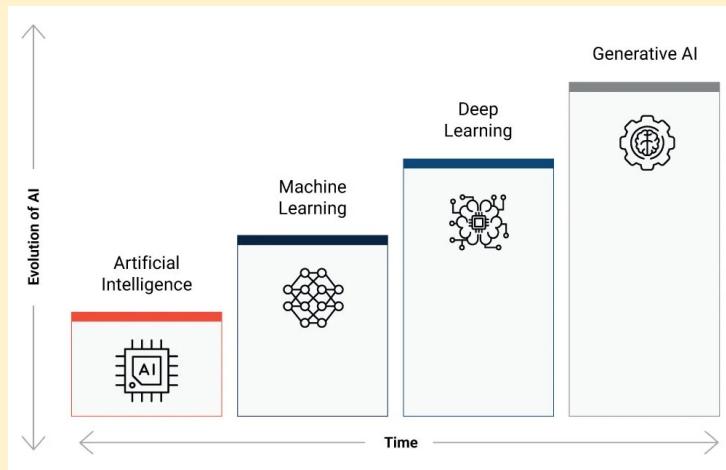
Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University

Peerapon.v@chula.ac.th



Outline

- Generative AI (Gen AI): LLM
- LLM Use Case
- Gen AI's Tools



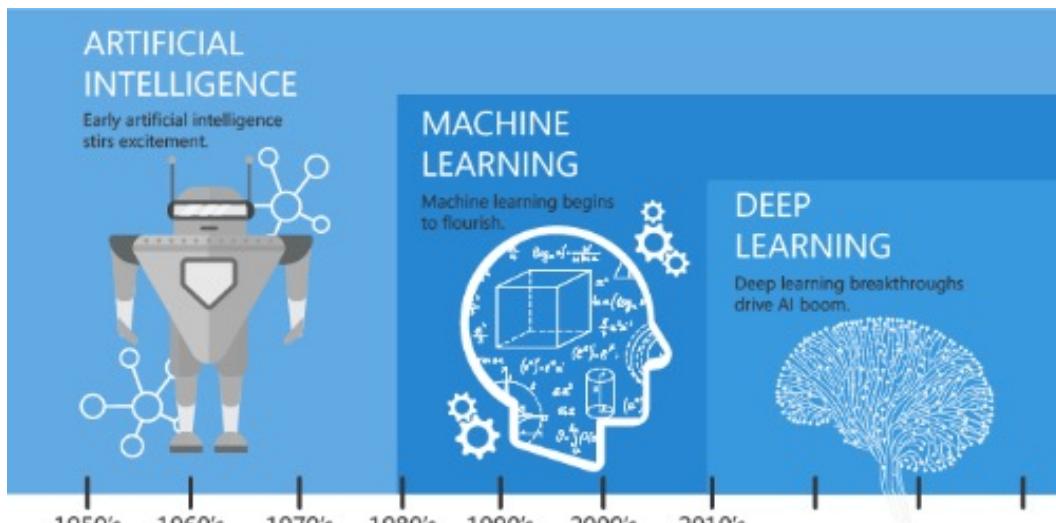
- Image (DALL E, Midjourney)
- Text (ChatGPT)
- Code (Codex)

Generative AI: LLM

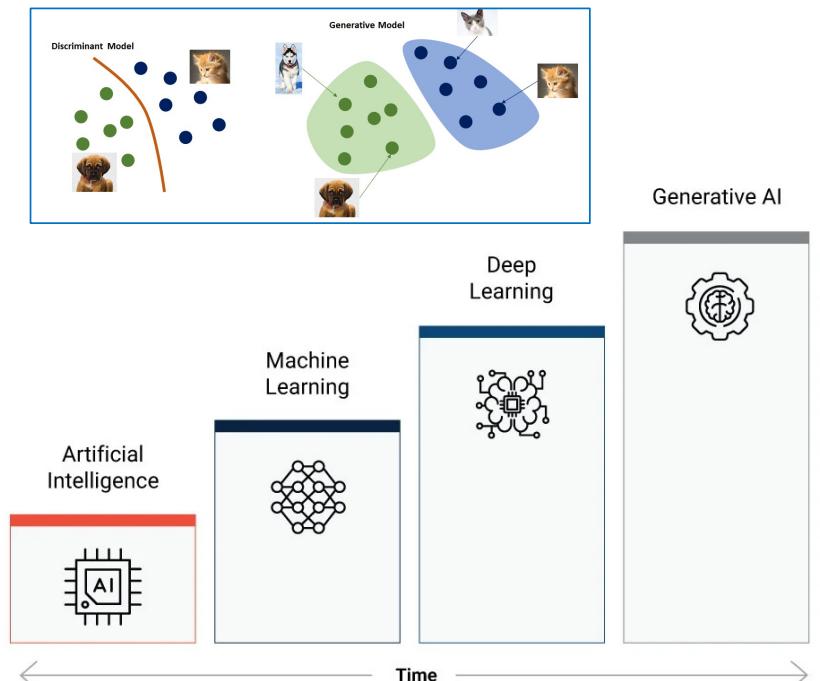
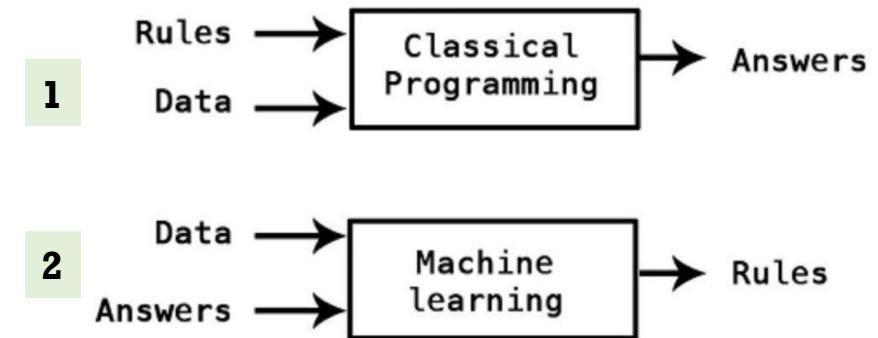
Credit to Aj.Ekapol's slide & Aj.Thanachart's slide (IMC)

AI = Automation

- 1) Rule-based AI
- 2) Machine Learning (ML)

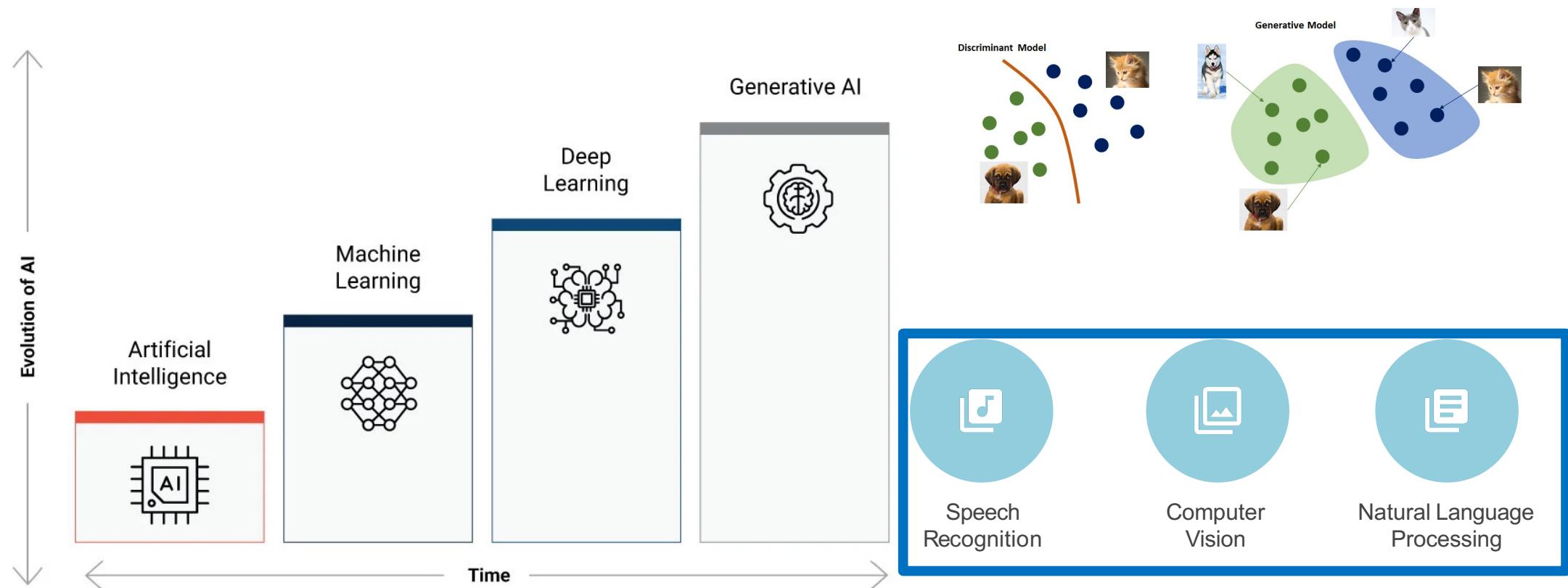


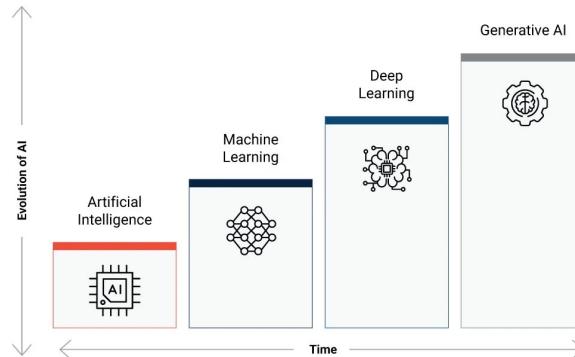
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



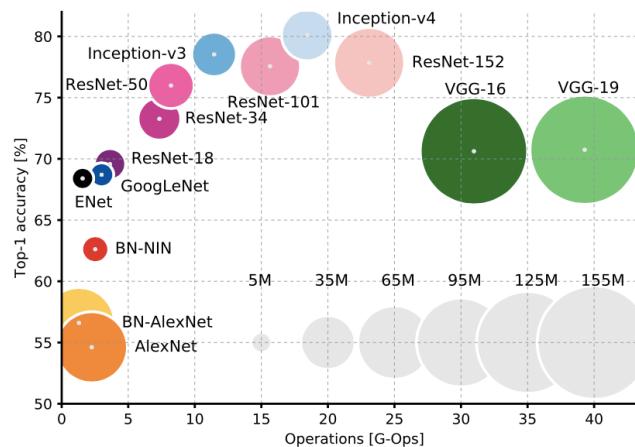
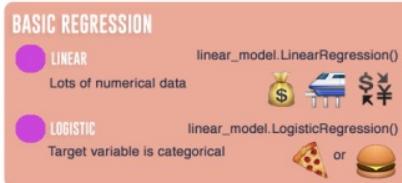
<https://mc.ai/machine-learning-basics-artificial-intelligence-machine-learning-and-deep-learning/>

Recent Research: DL & Generative AI

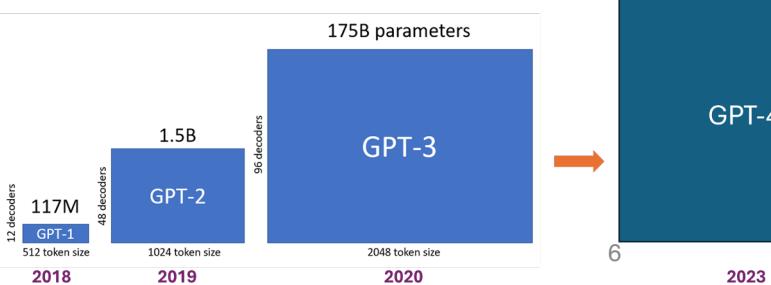




$$\text{Spend} = 500 + 10^* \text{Age} + 20^* \text{Income}_1K$$

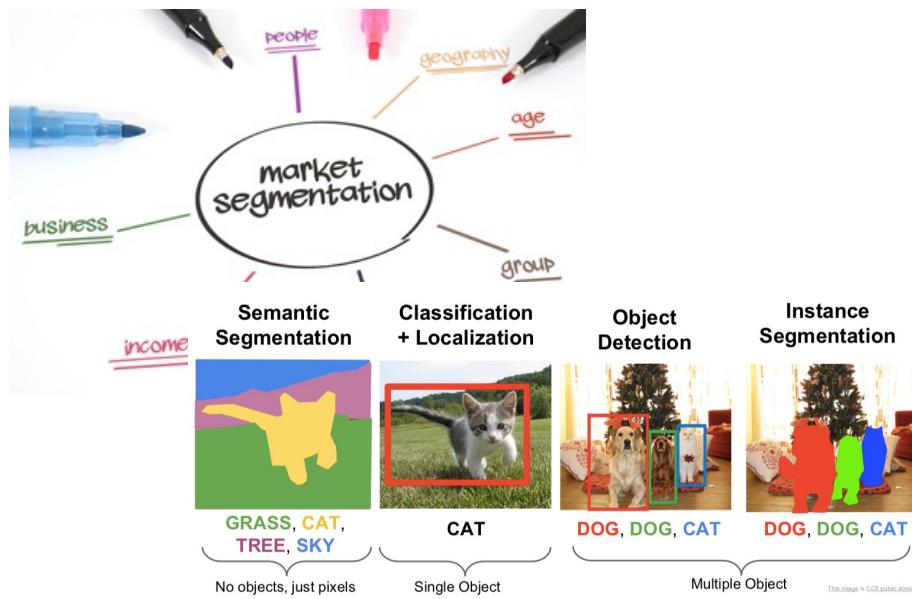


How exactly “Large” it is

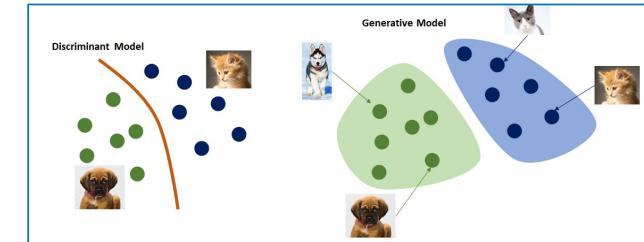


Predictive AI vs. Generative AI

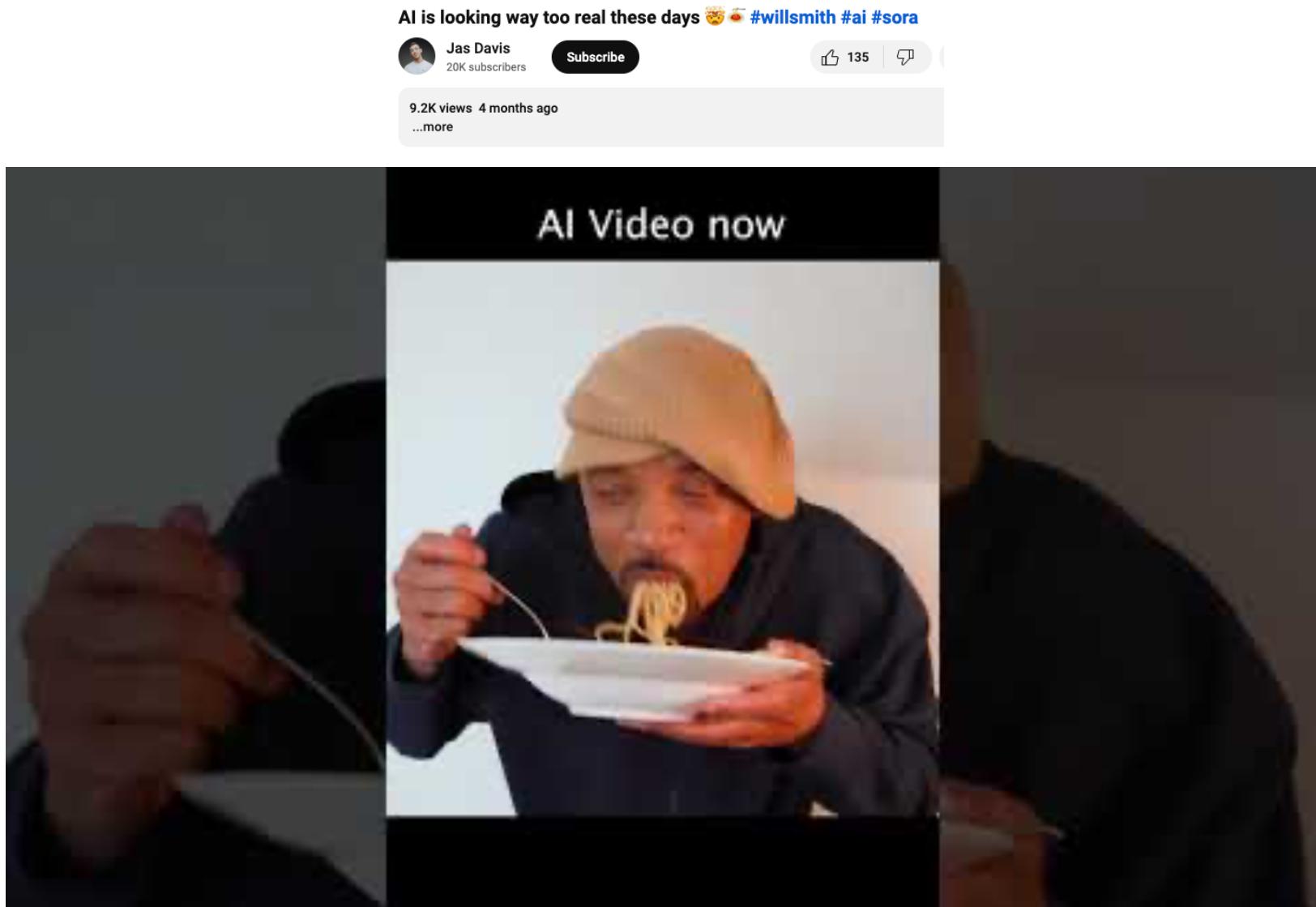
- Predictive AI
 - Predict, Forecast, Detect
 - Cluster, Classify, Segment



- Generative
 - Generate, Generalize



<https://www.youtube.com/watch?v=XQr4Xklqzw8>



<https://openai.com/index/sora/>

ChatGPT

- ChatGPT was launched by OpenAI on **30 Nov 2022**.
- ChatGPT is a **large language model (LLM)** for conversational AI applications.
- **Generates** human-like text and performs NLP tasks.
- Scalable and flexible for various use cases.
- **ChatGPT didn't enclose the details.**

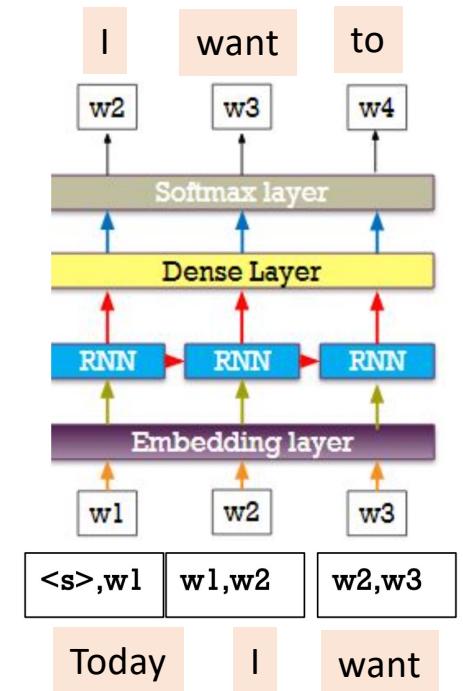


Language Model (LM)

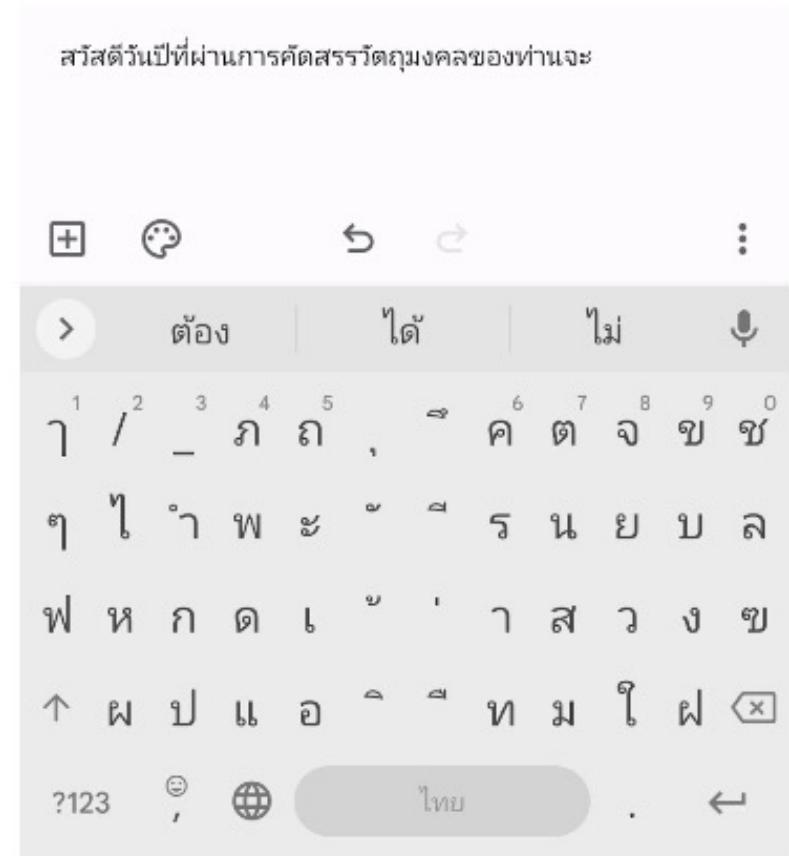
- It is the model that aims to predict next word based on the given previous words.
- So, the model can understand grammar & context.

วันนี้เรารอยากจะกินข้าวมัน _____

Today I want to eat chicken rice



Generate text with AI



Generate text with AI

2015

Obama-RNN [2015]

Data: 730,895 tokens (4MB)

Model: 3MB

Good afternoon. God bless you.

The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done. The promise of the men and women who were still going to take out the fact that the American people have fought to make sure that they have to be able to protect our part. It was a chance to stand together to completely look for the commitment to borrow from the American people. And the fact is the men and women in uniform and the millions of our country with the law system that we should be a strong stretches of the forces that we can afford to increase our spirit of the American people and the leadership of our country who are on the Internet of American lives.

Thank you very much. God bless you, and God bless the United States of America.

สร้างข้อความด้วย AI

2015

Obama-RNN [2015]

Data: 730,895 tokens (4MB)

Model: 3MB

Good afternoon. God bless you.

The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done. The promise of the men and women who were still going to take out the fact that the American people have fought to make sure that they have to be able to protect our part. It was a chance to stand together to completely look for the commitment to borrow from the American people. And the fact is the men and women in uniform and the millions of our country with the law system that we should be a strong stretches of the forces that we can afford to increase our spirit of the American people and the leadership of our country who are on the Internet of American lives.

Thank you very much. God bless you, and God bless the United States of America.

Top Large Language Models

Obama-RNN [2015]
Data: 730,895 tokens (4MB)
Model: 3MB parameters

2020



2021



LaMDA (173B)



MT-NLG (530B)

<https://vectara.com/top-large-language-models-langs-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other/>

GPT3 [2020]
Data: ~750GB (30,000x)
Model: 175B parameters (700,000x)
Training cost: \$5M, equivalent to ~300 years



PaLM (540B)



Flan-UL2 (20B)



GATO (1B)

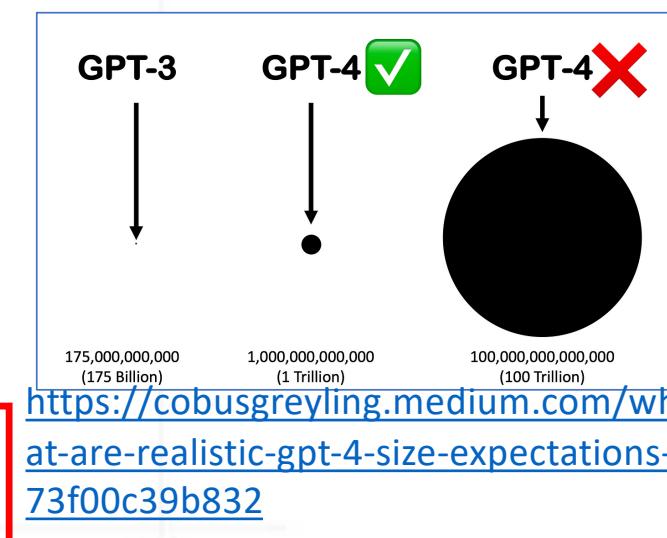
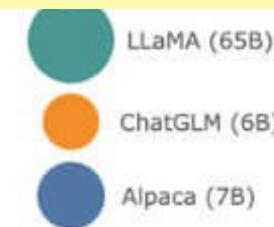


BLOOM (176B)



ChatGPT (20B)

2022





MARKETS TECHNOLOGY MONEY HEALTHCARE ENERGY MAPS MINING GREEN PARTNER WITH US

CREATOR PROGRAM

REPORTS

CHANNELS



TECHNOLOGY

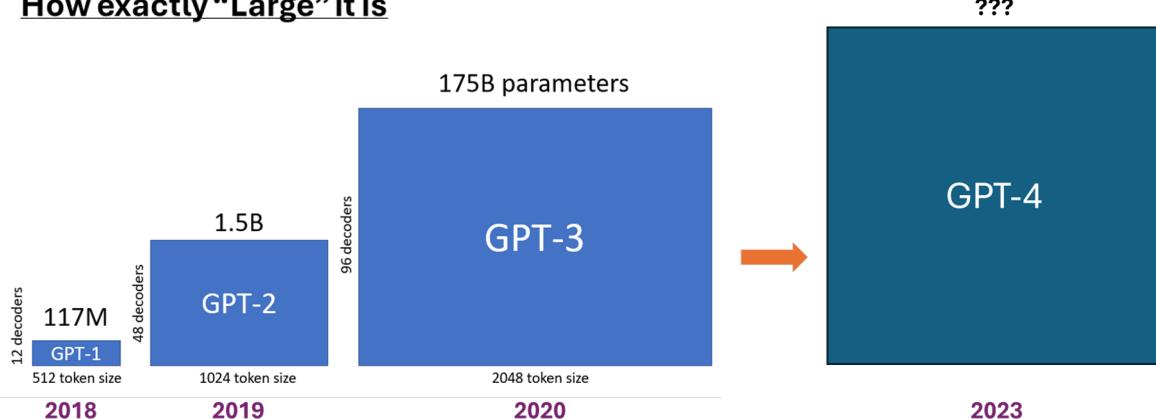
Visualizing the Training Costs of AI Models Over Time



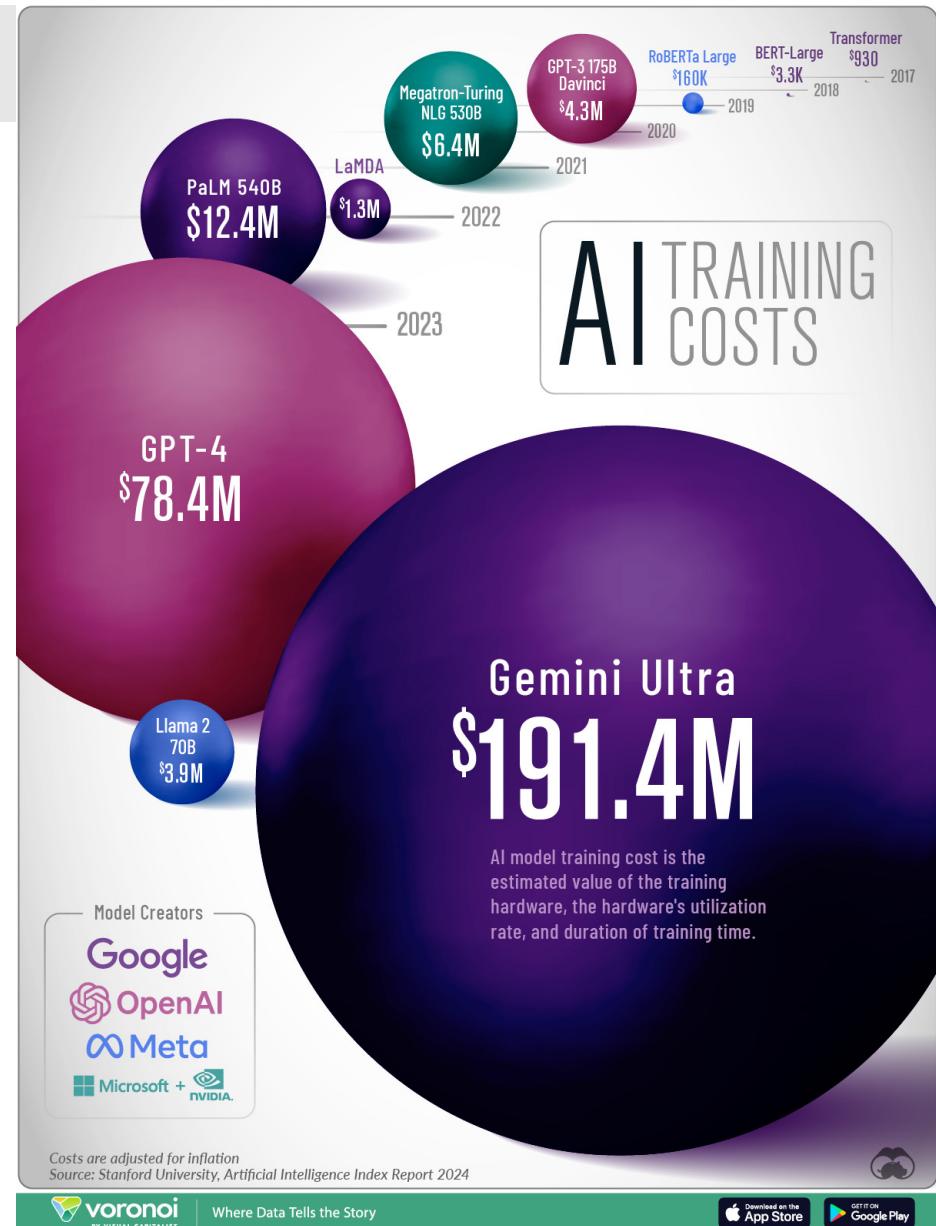
Published 3 months ago on June 4, 2024
By Dorothy Neufeld

- GPT-3: \$4.3M (144M THB)
- GPT-4: \$78.4M (2.6B THB)

How exactly “Large” it is



<https://www.visualcapitalist.com/training-costs-of-ai-models-over-time/>



OpenAI



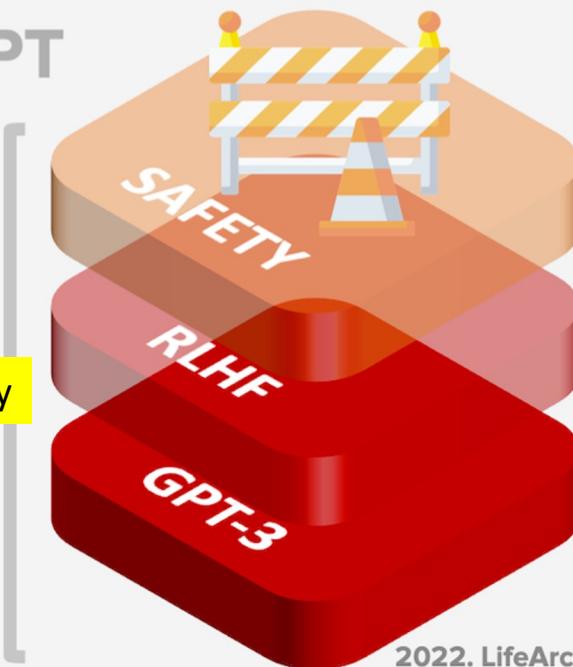
InstructGPT

Jan/2022



ChatGPT

Nov/2022



Step 1

Collect demonstration data and train a supervised policy.

LM

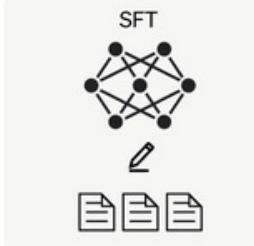
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

Reward Model (Scoring Model)

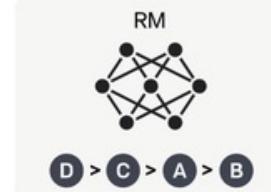
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

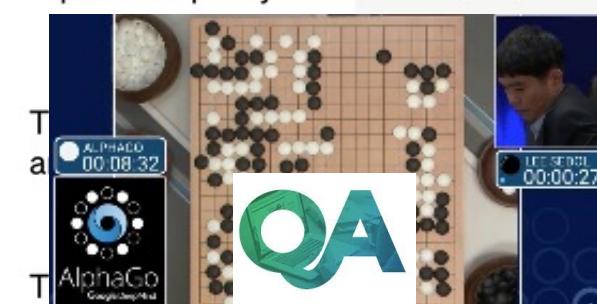
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

RL

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



QA calculates a reward for the output.

The reward is used to update the policy using PPO.

r_k

ChatGPT's capabilities

- Generate text
- Paraphrase
- Summarize

Context → Article:

Informal conversation is an important part of any business relationship. Before you start a discussion, however, make sure you understand which topics are suitable and which are considered taboo in a particular culture. Latin Americans enjoy sharing information about their local history, art and customs. You may expect questions about your family, and be sure to show pictures of your children. You may feel free to ask similar questions of your Latin American friends. The French think of conversation as an art form, and they enjoy the value of lively discussions as well as disagreements. For them, arguments can be interesting and they can cover pretty much or any topic ---- as long as they occur in a respectful and intelligent manner.

In the United States, business people like to discuss a wide range of topics, including opinions about work, family, hobbies, and politics. In Japan, China, and Korea, however, people are much more private. They do not share much about their thoughts, feelings, or emotions because they feel that doing so might take away from the harmonious business relationship they're trying to build. Middle Easterners are also private about their personal lives and family matters. It is considered rude, for example, to ask a businessman from Saudi Arabia about his wife or children.

As a general rule, it's best not to talk about politics or religion with your business friends. This can get you into trouble, even in the United States, where people hold different religious views. In addition, discussing one's salary is usually considered unsuitable. Sports is typically a friendly subject in most parts of the world, although be careful not to criticize national sport. Instead, be friendly and praise your host's team.

Q: What shouldn't you do when talking about sports with colleagues from another country?

A: Criticizing the sports of your colleagues' country.

Q: Which is typically a friendly topic in most places according to the author?

A: Sports.

Q: Why are people from Asia more private in their conversation with others?

A: They don't want to have their good relationship with others harmed by informal conversation.

Q: The author considers politics and religion . .

A:

Correct Answer → taboo

Prompt engineer aims to solve alignment problem.

- Language Model (LM) is originally trained to predict the next word, NOT answer the question.
- GPT (GPT3 is 175B parameters) is usually frozen (not trained).
- Since we cannot change the model, we need to align (change) the question (also called prompt).

| Input (Prompt) | Output |
|---|---|
| The patient was died. | The patient's body was found in a dark alley behind the hospital's... |
| "The patient was died." correct this | claim if you really believe such figures.... |
| Poor English input: The patient was died. | Good English output: The patient died. |



Jobs of the Future: AI Prompt Engineer



Cody W Burns

Emerging Technology Visionary | Distributed Systems | Privacy | Executive Leadership

10 articles

+ Follow

October 19, 2022

JOB OF THE FUTURE: AI PROMPT ENGINEER

Cody Burns
20

<https://www.linkedin.com/pulse/jobs-future-ai-prompt-engineer-cody-w-burns/>

<https://fireflies.ai/blog/ai-text-generators>

AI text generation tools

The image displays three AI text generation tools side-by-side:

- ChatGPT**: A white interface with a sidebar showing recent activity like "New chat", "ChatGPT 3.5", and "Previous 30 Days". The main area features a large "ChatGPT" logo, a small AI icon, and the question "How can I help you today?". Below it are two input fields: "Write a spreadsheet formula to convert a date to the weekday" and "Plan an itinerary to experience the wildlife in the Australian outback".
- Google Gemini**: A white interface with a sidebar showing "March" and "Gemini". The main area features a large "Google Gemini" logo, a small AI icon, and the greeting "Hello, Ayush.". Below it is the question "How can I help you today?". At the bottom, there are four suggested prompts: "Help me compare these college majors", "Help me get organized with a list of 10 tips", "Ideas to surprise a friend on their birthday", and "Help design a database schema for a business".
- Anthropic Claude**: A light gray interface with a sidebar showing "Hey" and "AK". The main area features a large "Anthropic Claude" logo, a small AI icon, and the question "What can I help you with?". Below it is the brand name "Claude by ANTHROPIC". At the bottom, there is a message input field with the placeholder "Message Claude..." and a red send button.



May 13, 2024

Hello GPT-4o

We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

[Contributions](#) > [Try on ChatGPT ↗](#) [Try in Playground ↗](#) [Rewatch live demos >](#)

GPT-4o ("o" for "omni") is a step towards much more natural human-computer interaction—it accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds, which is similar to human response time in a conversation. It matches GPT-4 Turbo performance on text in English and code, with significant improvement on text in non-English languages, while also being much faster and 50% cheaper in the API. GPT-4o is especially better at vision and audio understanding compared to existing models.

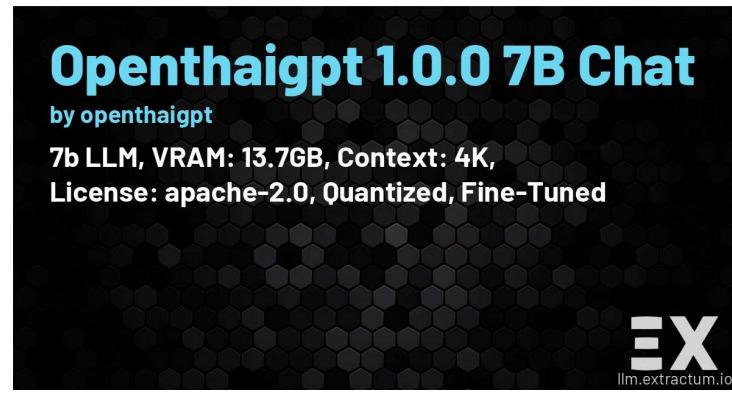
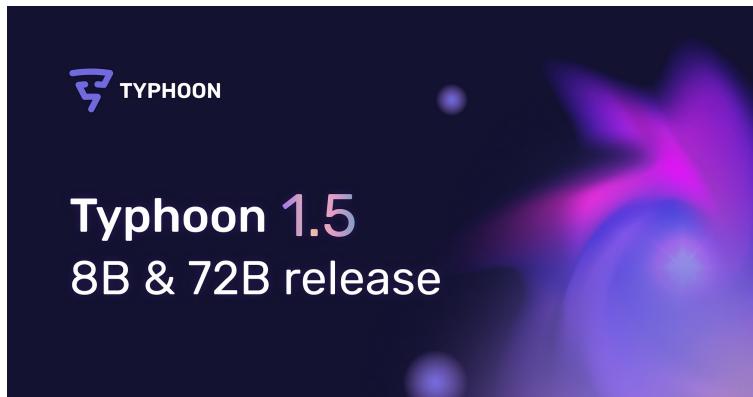
<https://openai.com/index/hello-gpt-4o/>

Commercial Large Language Model (LLM)

Global LLMs: OpenAI GPT, Google Gemini, LLaMA etc.

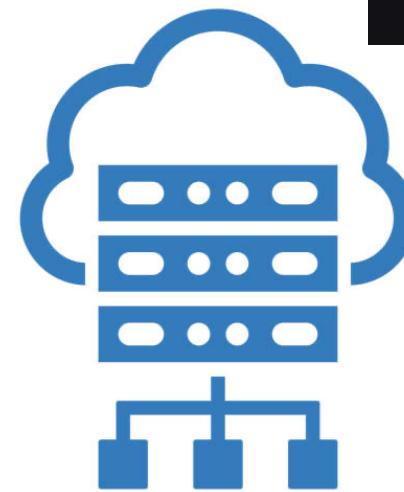
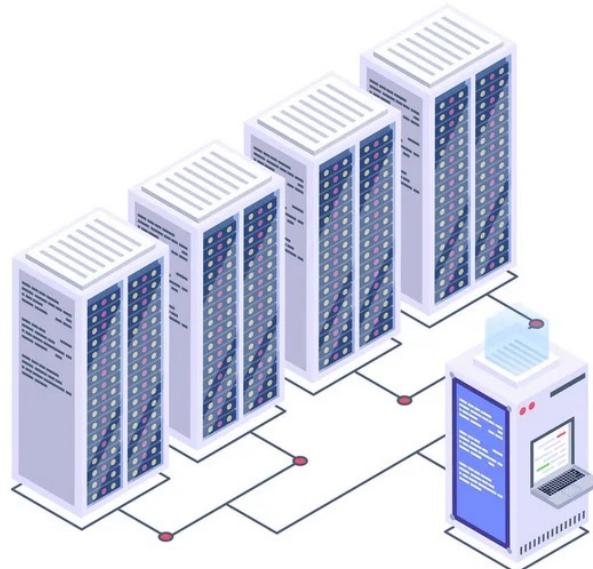
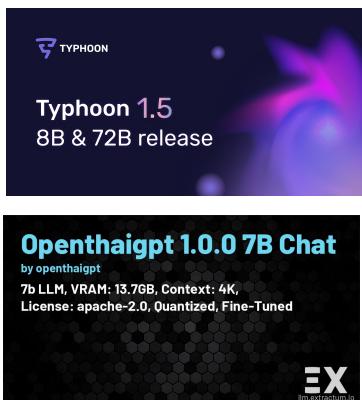


Thai LLMs: OpenThaiGPT, Typhoon, etc.



Loading a Pretrained Model Locally vs. Using a Cloud-based API LLM

LLaMA
by  Meta

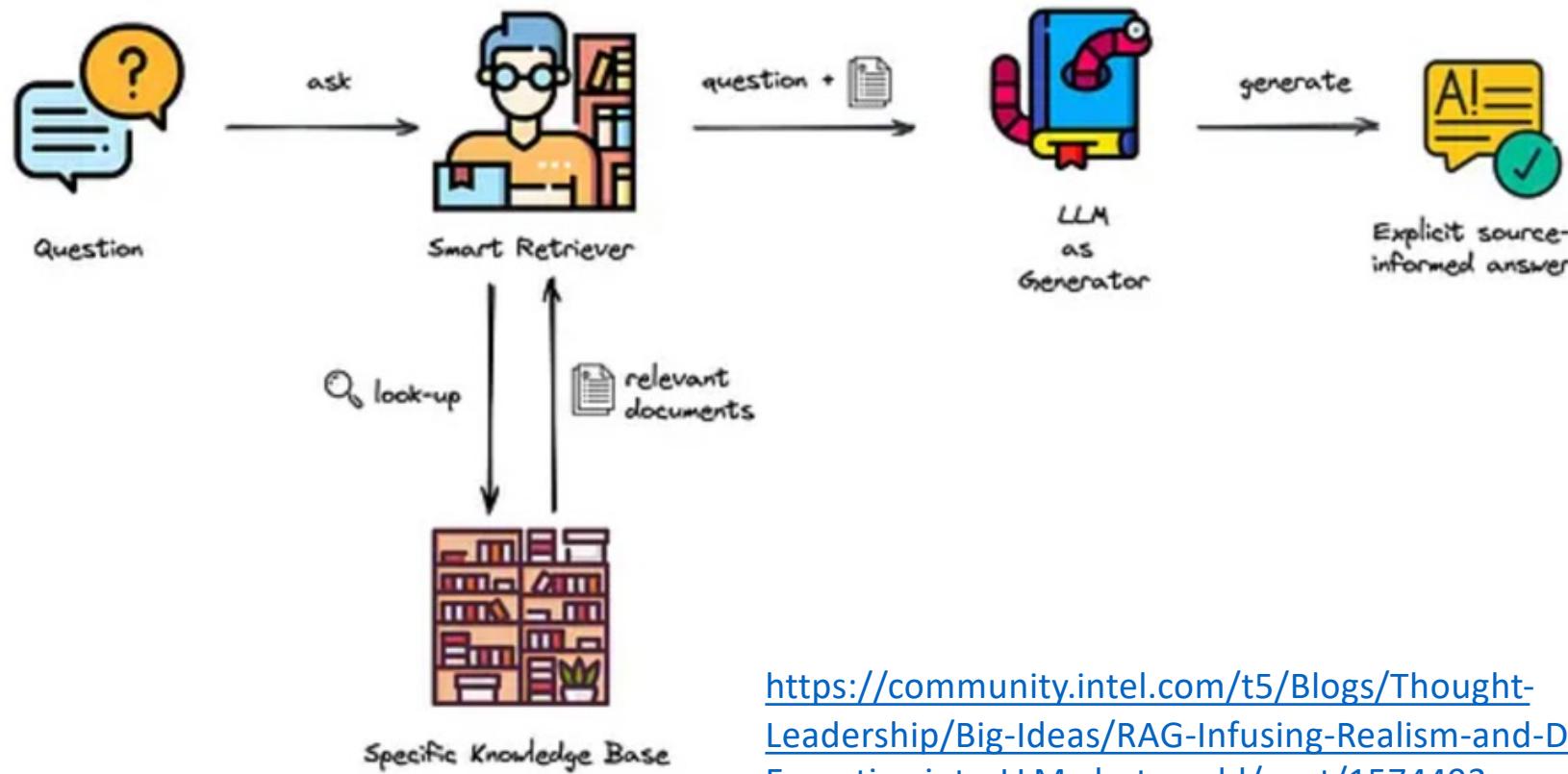


LLM's use cases

Customer service (QA & Chatbot)

```
1 delimiter = "####"
2 system_message = f"""
3 You are a customer service assistant at a retail company. \
4 You will be provided with customer service queries. \
5 The customer service query will be delimited with {delimiter} characters.
6 Hierarchical Classify each query into a primary category and a secondary category.
7 All secondary categories must belong to primary categories.
8 You will get short explanations for each categories inside && characters. \
9 Importantly, all the category you answer must strictly satisfy the short explanations if given.\n10 Provide your output in json format with the keys: primary and secondary. \
11
12 Primary categories:
13 - Shopping
14 - Returns and Payments
15 - Promotions and Discounts
16 - Products
17 - General Inquiry
18
19 if "Shopping" is a primary category, secondary categories must follow:
20 - Order Status and Tracking
21 - Compare Products
22 - Delivery Service
23 - Missing or Damaged Items in orders
24 - Buy Products
25
26 if "Returns and Payments" is a primary category, secondary categories must follow:
27 - Returns and Refunds
28 - Bill
29
30 if "P
31 - Ong
32 - App
33 - Cla
34
35 if "P
36 - Pro
37 - Pro
38 - Siz
39 - War
40
41 if "G
42 - Sto
43 - Pro
44 - Com
45 - Spe
46
47 """
query: 1
user query : I placed an order a few days ago, and I'm wondering when it will be delivered. Can you provide me with the shipping status?
output : {"primary": "Shopping", "secondary": "Order Status and Tracking"}
token : {'prompt_tokens': 362, 'completion_tokens': 19, 'total_tokens': 381}
-----
query: 2
user query : I received a damaged item in my order. What should I do to get a replacement or a refund?
output : {"primary": "Returns and Payments", "secondary": "Returns and Refunds"}
token : {'prompt_tokens': 355, 'completion_tokens': 21, 'total_tokens': 376}
-----
query: 3
user query : I'm interested in the new line of smartphones you have. Can you tell me more about their features and pricing?
output : {"primary": "Products", "secondary": "Product Information"}
token : {'prompt_tokens': 357, 'completion_tokens': 17, 'total_tokens': 374}
-----
query: 4
user query : Do you have any ongoing promotions or discount codes I can use for my next purchase?
output : {"primary": "Promotions and Discounts", "secondary": "Ongoing Sales or Promotions"}
token : {'prompt_tokens': 351, 'completion_tokens': 24, 'total_tokens': 375}
-----
query: 5
user query : I forgot my password and can't access my account. Can you help me reset it?
output : {"primary": "General Inquiry", "secondary": "Speak to a human"}
```

Retrieval Augmented Generation (RAG)

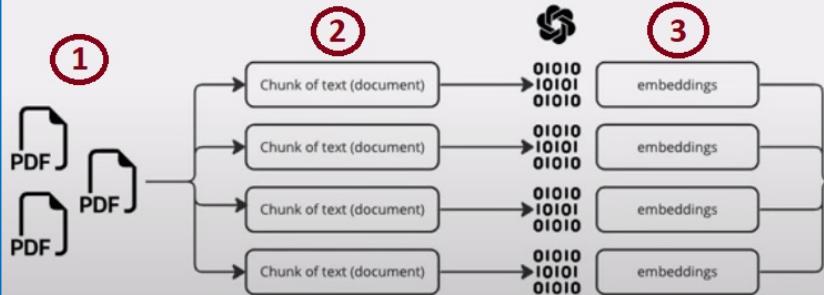
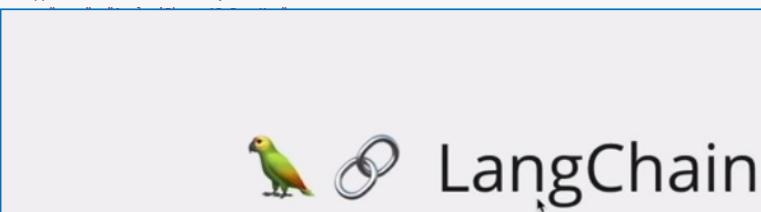


<https://community.intel.com/t5/Blogs/Thought-Leadership/Big-Ideas/RAG-Infusing-Realism-and-Domain-Expertise-into-LLMs-but-could/post/1574493>

RAG Tools

1. Product Information

```
1 # product information
2 products = [
3     "OPPO A58": {
4         "name": "OPPO A58",
5         "category": "Smartphones",
6         "brand": "OPPO",
7         "warranty": "24 months",
8         "features": ["6.72 in. LCD", "50 MP camera", "Octa Core 2.0 GHz", "RAM 6 GB", "ROM 128 GB", "battery 5,000 mAh 33 W"],
9         "description": "Simply stylish, this phone shimmers and shines with a luxurious silk-satin texture \\ from the innovative OPPO Glow design. Elegant curves fit your grip like a glove.",
10        "price": 182.00
11    },
12    "vivo V29": {
13        "name": "vivo V29",
14        "category": "Smartphones",
15        "brand": "vivo",
16        "warranty": "24 months",
17        "features": ["6.78 in. AMOLED", "50 MP camera", "Octa Core 2.4 GHz", "RAM 12 GB", "ROM 256/512 GB", "battery 4,600 mAh 80 W"],
18        "description": "Top-tier screen and design, innovative 3D magnetic particle. Fantastic design, \\ fascinating aura. Ignite every moment with delightful portraits.",
19        "price": 414.00
20    },
21    "Apple iPhone 15 Pro Max": {
22        ...
23    }
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51 }
```

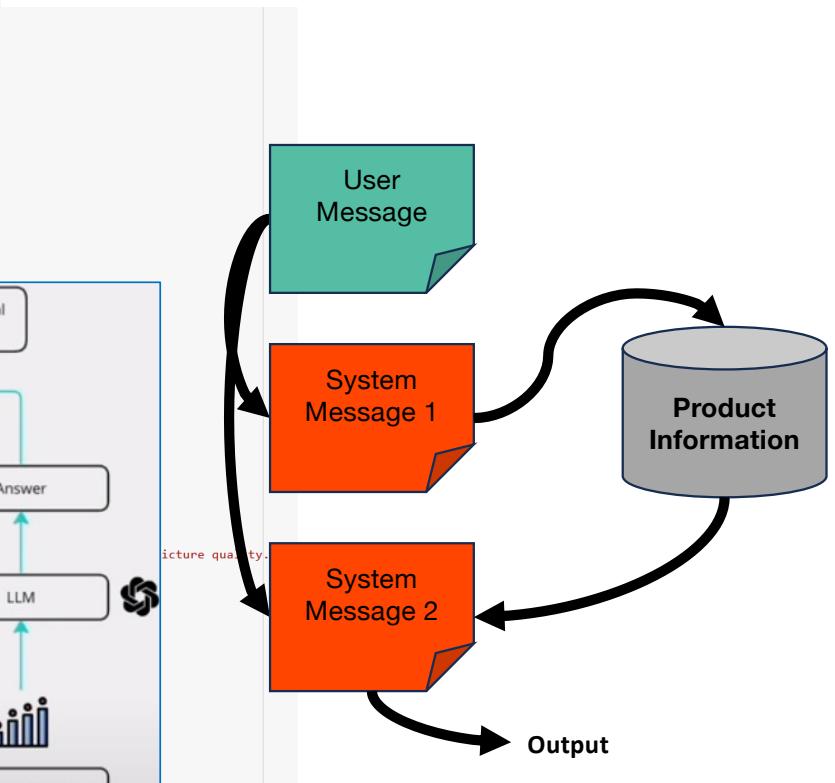


RAG Tools

Some popular comprehensive tools to build RAG systems include [LangChain](#), [LlamaIndex](#), and [DSPy](#).

There are also a range of specialized tools that serve different purposes such as [Flowise AI](#) that offers a low-code solution for building RAG applications. Other notable technologies include [HayStack](#), [Meltano](#), [Cohere Coral](#), and others. Software and cloud service providers are also including RAG-centric services.

For instance, Verba from Weaviate is useful for building personal assistant applications and Amazon's Kendra offers intelligent enterprise search services.



<https://www.promptingguide.ai/research/rag>

THAILAND / TECH

AI จะเป็น Game Changer ของวงการแพทย์ไทยหรือไม่ คุยกับ นพ.ปิยะ ฤทธิ์ อิทธิชัยวงศ์ หนอผู้คิดค้น AI เครื่องมือแพทย์เพื่อคนไทย

โดย กักรสุดา บุญญาศรี
26.09.2023



Overview

Our Goals

Partners

Explore

Try it

MEDICAL FOCUS GENERATIVE AI

PreceptorAI is your pocket 2nd opinion for medical professionals finely tuned to Thai clinical guidelines.

Try it now



iApp

Main Site AI API Docs Support Contact Us

E-Voting E-KYC Chochae GPT Robot **Thanoy** Blog Certification Join Our Team

Thanoy

กนอย

ผู้ช่วยกฎหมาย AI Chatbot

พนักงาน "กนอย" ขออภัยเดี๋วนี้สูบบุหรี่ Ai จึงเบี่ยง
ด้วยเพลิงของ Chochae GPT

Add friend

LINE QR code

<https://iapp.co.th/thanoy>

04:21

< 99+ หน่อย

ช่วยวิเคราะห์รูปคดี ตำรวจจับ 'ลุงเปี้ยก' จัดหลักฐานให้ 'ฆ่าเมีย' เตรียมส่งข้อการอ่าน 'นิยายสอนสวน' ลั่งฟื้องคดี มี 'แผนประทุยกรรม' ประกอบเห็นมื่นเป็นปี!
ใช้เก้าอี้ตีจนตาย แต่คนร้ายตัวจริงกลับไม่ใช่ ใจร้ายจัดให้ 'ลุงเปี้ยก' รับสารภาพ 'ข้อหาฆ่า' แทน 'แก้กลุกตำรวจ'?

10:02

คันพับนบทความดังต่อไปนี้

- ✓ ประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 420
- ✓ ประมวลกฎหมายอาญา มาตรา 157
- ✓ ประมวลกฎหมายอาญา มาตรา 210
- ✓ ประมวลกฎหมายอาญา มาตรา 310

จากการนี้ข่าวที่ตำรวจจับกุมตัว "ลุงเปี้ยก" และจัดจากให้รับสารภาพว่า เป็นผู้ฆ่าการพยายามตอนน่อง โดยใช้เก้าอี้ตีจนเดือด แต่จากการสืบสวน สอนสวนในภายหลังกลับพบว่า คนร้ายตัวจริงไม่ใช่ "ลุงเปี้ยก" แต่เป็นแก๊งลูกตำรวจ ซึ่งมีการวางแผนประทุยกรรมให้ "ลุงเปี้ยก" รับสารภาพแทน

ในการนี้ "ลุงเปี้ยก" อาจมีสิทธิ์ฟ้องร้องดำเนินคดีกับเจ้าหน้าที่ตำรวจที่เกี่ยวข้องในความผิดฐาน

- ปฏิบัติหน้าที่โดยมิชอบ ตามประมวลกฎหมายอาญา มาตรา 157 ซึ่งมีโทษจำคุกไม่เกิน 5 ปี หรือปรับไม่เกิน 100,000 บาท หรือทั้งจำทั้งปรับ

...

เขียน ▲

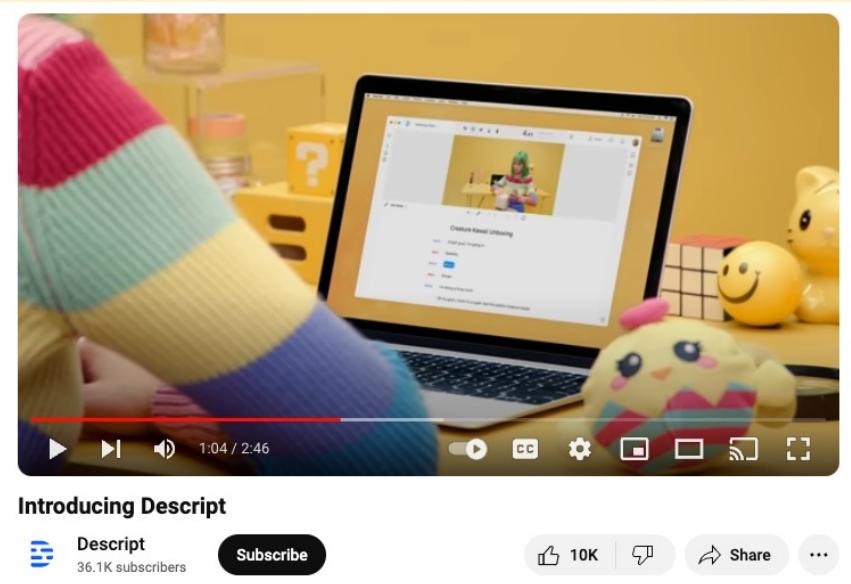
Use Cases in Agriculture



Reference: <https://www.biotec.or.th/home/biotec-launch-dapbot-th/>

Gen AI's Tools

- Text → Text
- Text → Image
- Text → Video



<https://www.youtube.com/watch?v=Bl9wqNe5J8U>

- Image → Text (captioning)
- Video → Text (captioning)

- Speech → Text
- Text → Speech

The screenshot shows the AI Marketplace interface. At the top, there are tabs for Dashboard, AI Marketplace (which is selected), and Library. On the right, there are buttons for Sign in, API Document, and Add to Library. Below the tabs, it says "Speech To Text". It displays metrics: WER: 10.6%, CER: 4.7%, DER: 13.5%, and TH. A note states: "AI can transcribe speech into text, supporting multiple file types such as mp3, wav, and flac." It credits "Publisher: Gowajee". There are three main buttons at the bottom: "Let's try demo", "How to call the API?", and "Full evaluation". The "Let's try demo" button is highlighted. To the right, there is a section for "Uploaded file" with a placeholder box.

<https://console.visai.ai/ai-marketplace/ready-to-use/c7ba44f3-0de6-4457-8743-75438e3eac60/speech-to-text>

<https://fireflies.ai/blog/ai-text-generators>

Text → Text (QA)

The screenshot shows the Fireflies AI platform interface. On the left, there's a sidebar with various AI-related links and a main chat area for ChatGPT 3.5. In the center, there's a large input field with the placeholder "What can I help you with?". Below it, the Anthropic Claude model is active, with its logo and name in a blue-bordered box. At the bottom, there's a message input field with "Message Claude..." and a send button.

AI and ML

Top 10 AI Text Generators [Free & Paid]



Ayush Kudesia
Apr 27, 2024

The screenshot shows a blog post titled "Top 10 AI Text Generators [Free & Paid]" by Ayush Kudesia. The post features a section for ChatGPT, another for Anthropic Claude, and a larger section for Google Gemini. The Gemini section includes a "Hello, Ayush." greeting, a "How can I help you today?" prompt, and a row of four cards with prompts: "Help me compare these college majors", "Help me get organized with a list of 10 tips", "Ideas to surprise a friend on their birthday", and "Help design a database schema for a business". There's also a "Enter a prompt here" input field at the bottom.

[Best apps](#) 17 min read

The best AI image generators in 2024

Get the best AI-generated images using text-to-image AI.

By Harry Guinness · February 22, 2024

- [DALL-E 3](#)
- [Midjourney](#)
- [Stable Diffusion](#)
- [Adobe firefly](#)
- [ImageFX by Google](#)



- [Synthesia](#)
- [Fliki](#)
- [Peech](#)
- [Colossyan](#)
- [Runway](#)

[Best apps](#) 15 min read

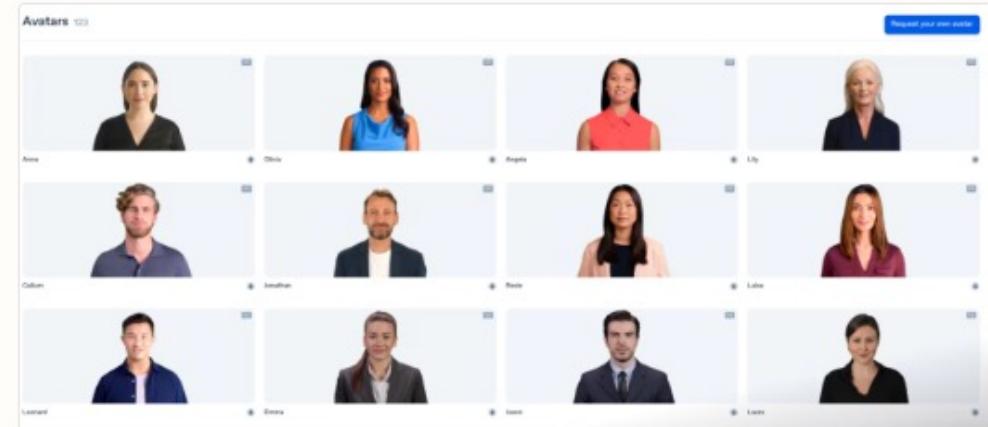
The best AI video generators

Use these tools to create, edit, and enhance videos with AI.

By Miguel Rebelo · August 2, 2023

Best AI video generator for using digital avatars

[Synthesia \(Web\)](#)



Video Captioning: NoteGPT

YouTube Video Summary Generator Free Online with AI

Summarize YouTube Videos with one click

<https://www.youtube.com/watch?v=2njn71TqkjA>

Generate Summary

YouTube Video

Transcript Summary MindMap AI Chat Save as Note

Translate

Summary

In 2050, Earth will face devastating effects of climate change if current trends continue, with rising temperatures leading to heatwaves, droughts, and sea level rise. However, there is still time to mitigate these impacts through bold actions and collective efforts.

Highlights

- ● Earth in 2050 faces a grim future with 2 degrees temperature rise and frequent extreme heatwaves and wildfires.
- ● Rising temperatures lead to power outages, healthcare crises, and water scarcity in many regions.
- ● Entire nations become uninhabitable due to sea level rise, displacing millions of climate migrants.
- ● Food scarcity and price hikes affect global communities, leading to hunger and famine.
- ■ Bold policies and innovations are needed to combat climate change and mitigate its worst impacts.
- ● Collective action is crucial to rewrite our future and prevent further climate devastation.
- ● Time is running out to make meaningful changes and secure a sustainable

What Earth in 2050 could look like - Shannon Odell

Author TED-Ed

Video Duration 05 mins 00 secs

Read Time 1 min 5 secs

Write a feedback

Example: What Earth in 2050 Could Look Like

