



Large Language Model (LLM)

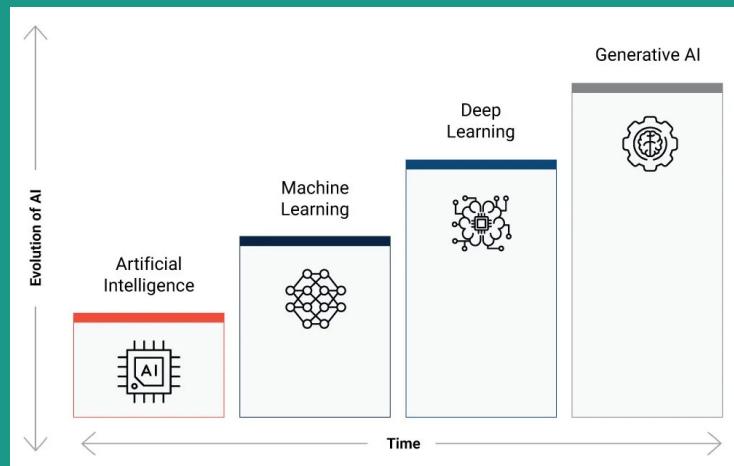
Associate Professor Peerapon Vateekul, Ph.D.

peerapon.v@chula.ac.th

Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University

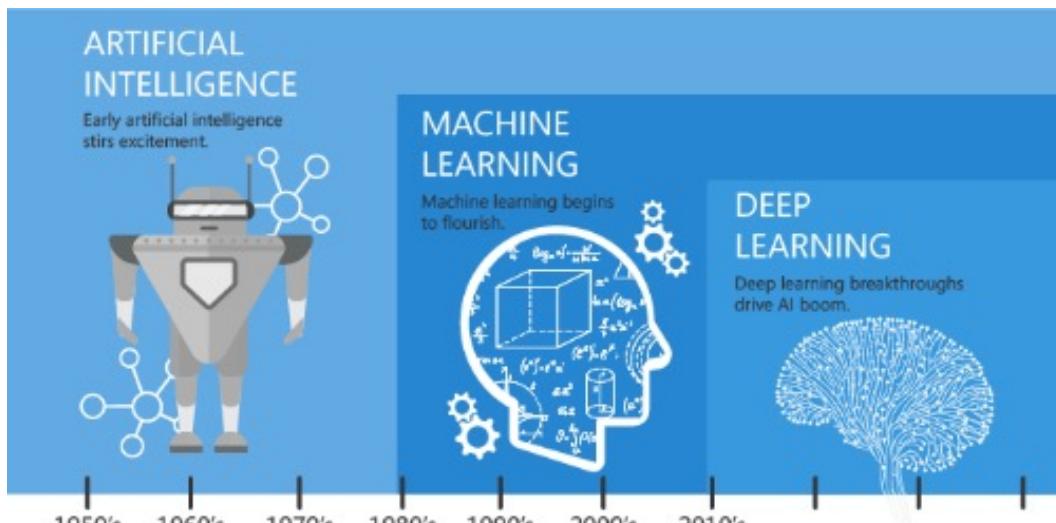
Generative AI: LLM

Credit to Asst. Prof.Ekapol's slide

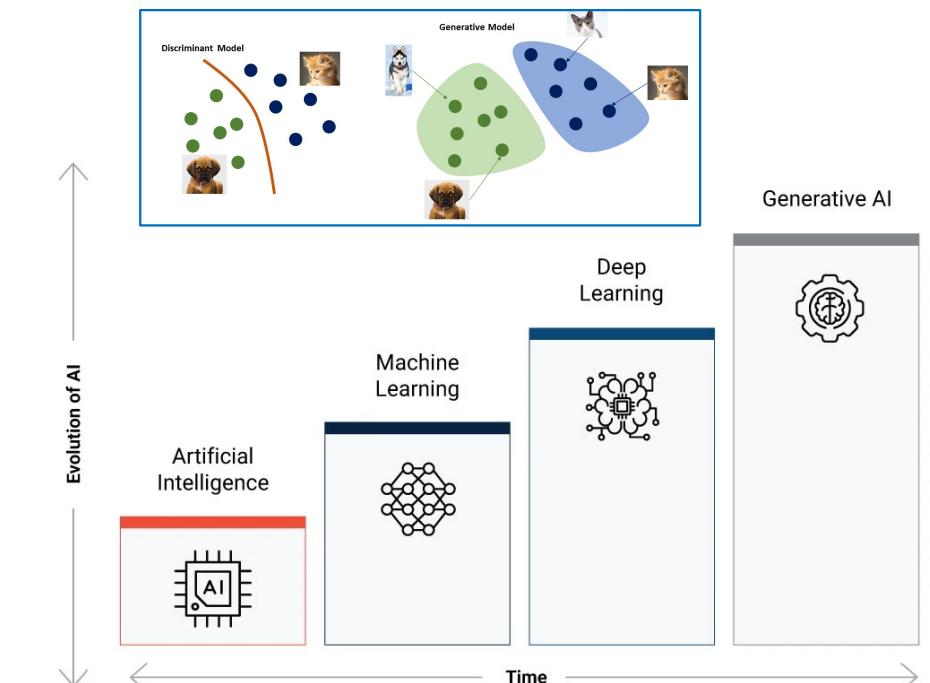
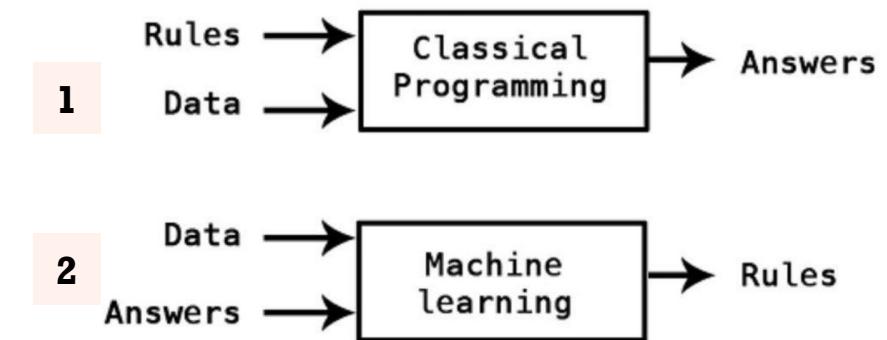


AI = Automation

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

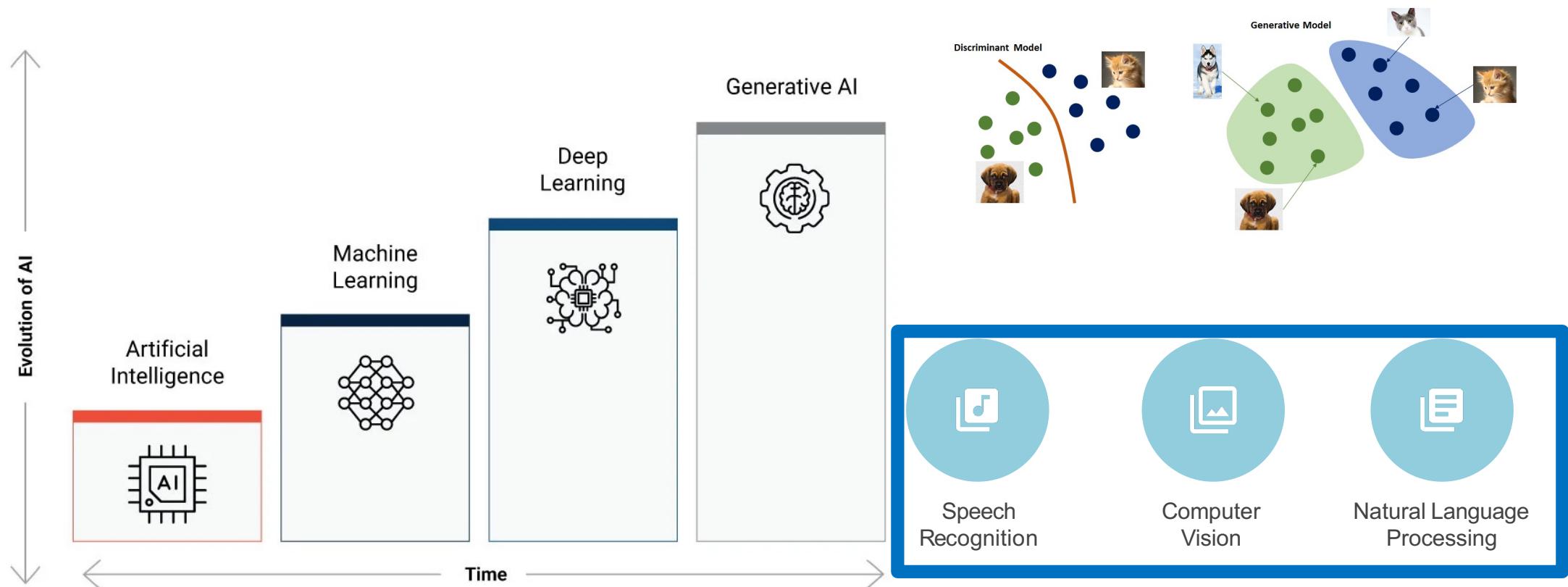


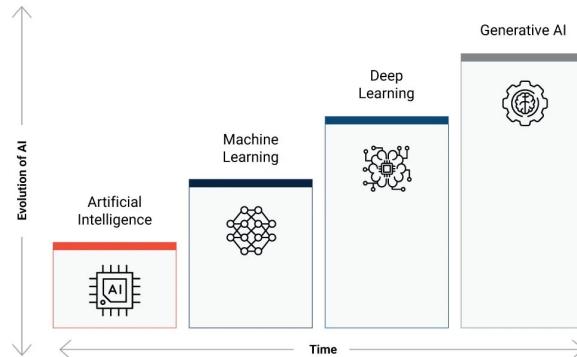
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



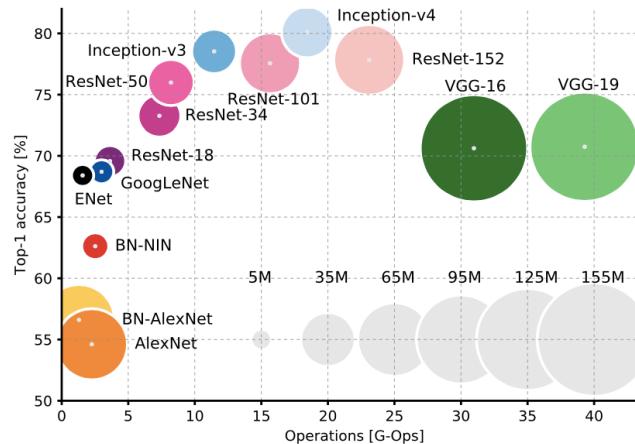
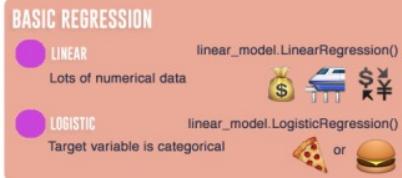
<https://mc.ai/machine-learning-basics-artificial-intelligence-machine-learning-and-deep-learning/>

Recent Research: DL & Generative AI

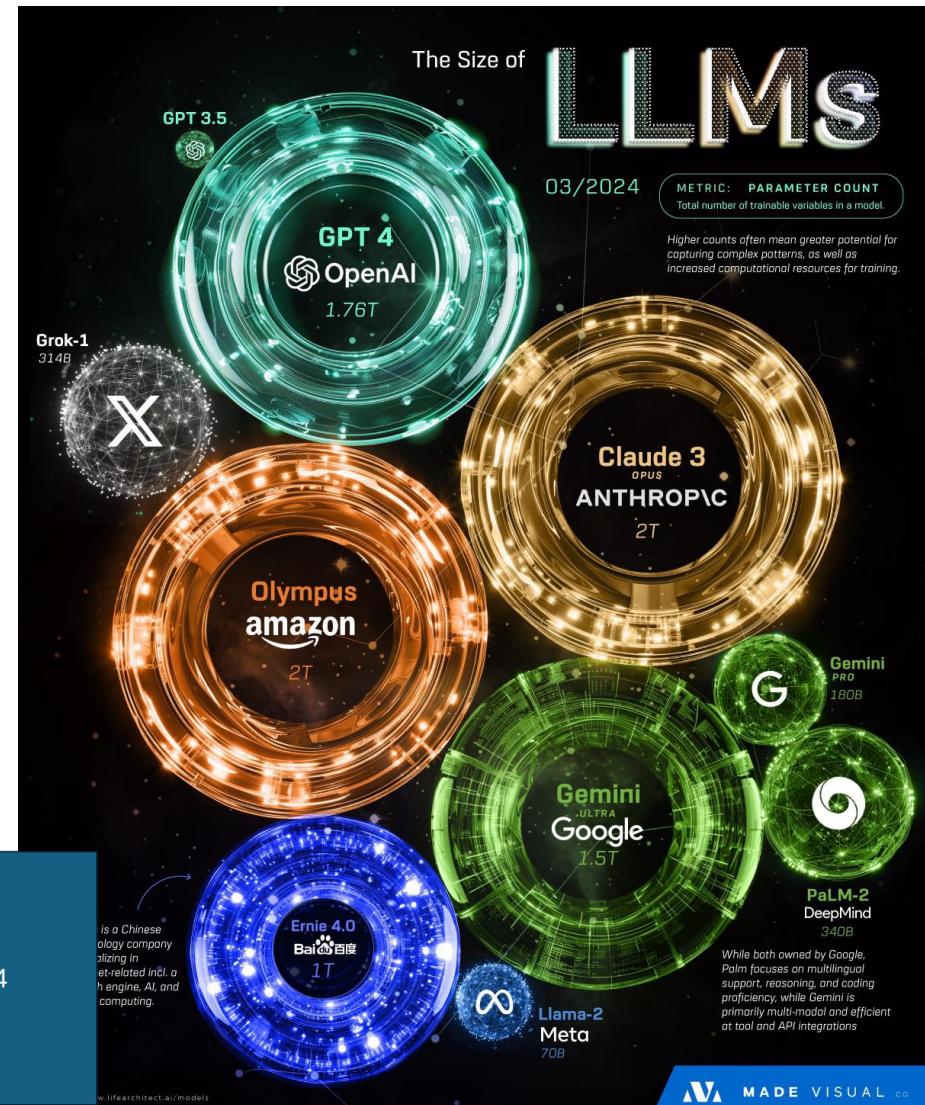
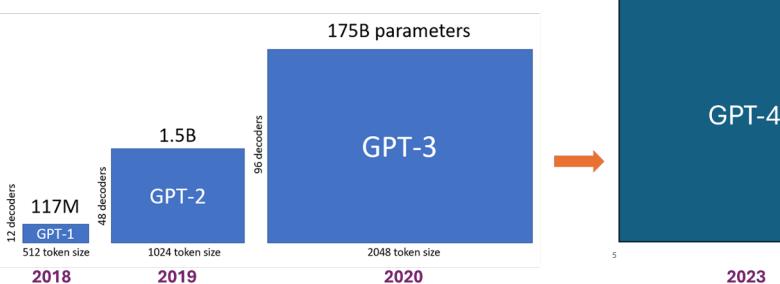




$$\text{Spend} = 500 + 10 * \text{Age} + 20 * \text{Income}_1K$$

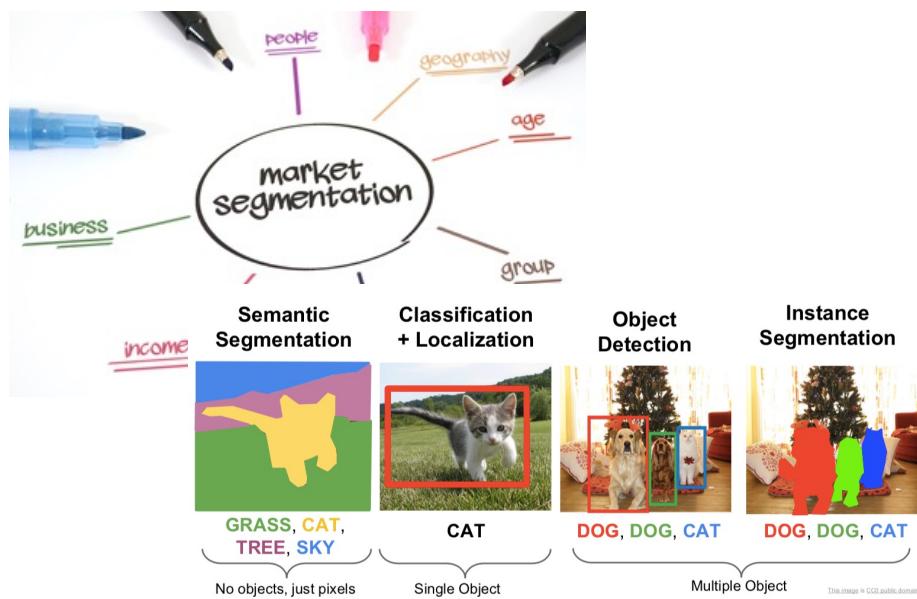


How exactly “Large” it is

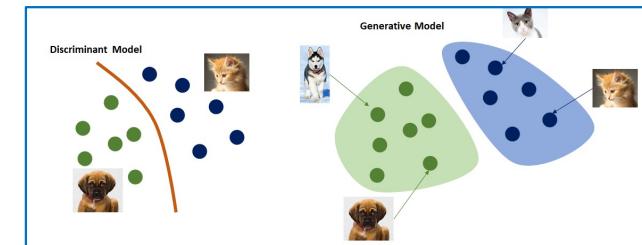


Predictive AI vs. Generative AI

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



- Generative
 - Generate, Generalize



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

AI is looking way too real these days 😱🔥 #willsmith #ai #sora



Jas Davis
20K subscribers

Subscribe

135



9.2K views 4 months ago
...more

AI Video now



<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.



OpenAI

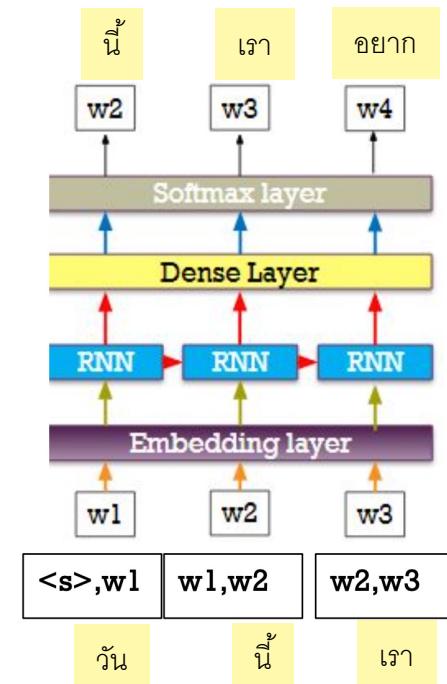


Language Model (LM)

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

วันนี่เราอยากจะกินข้าวมัน _____

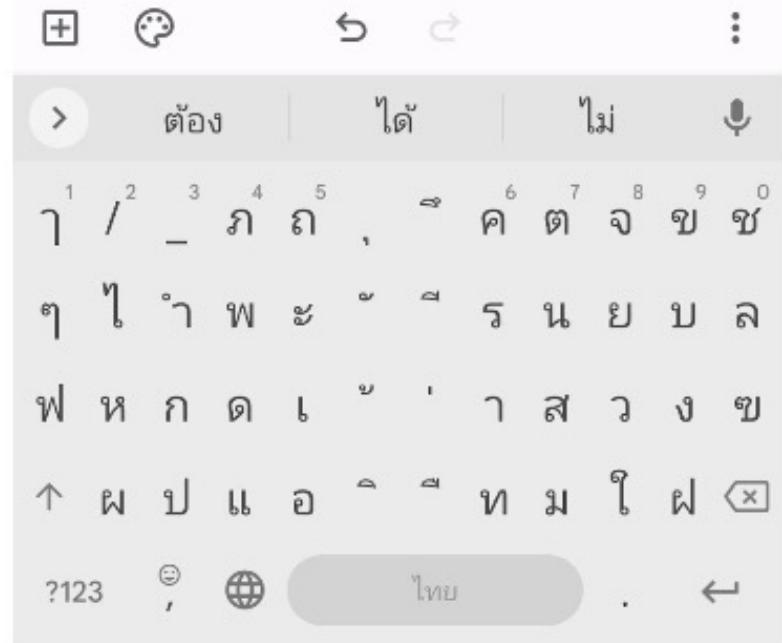
วัน นี่ เรา อยาก จะ กิน ข้าว มัน



สร้างข้อความด้วย AI



สวัสดีวันเป็นผ่านการคัดสรรร่วมกับของท่านจะ



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



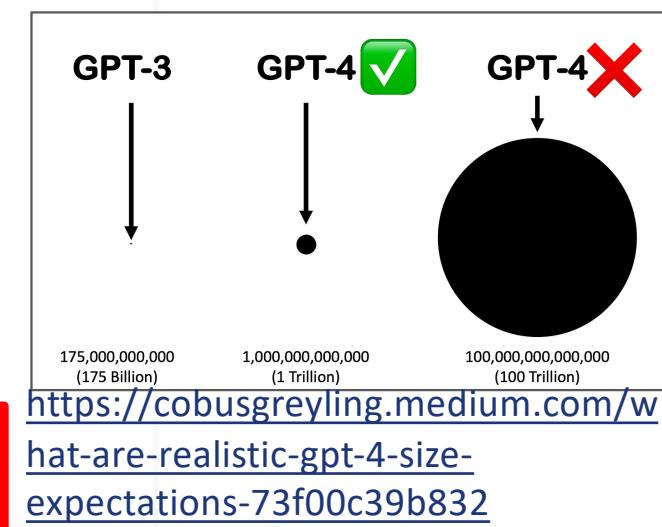
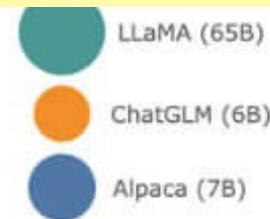
2021



ChatGPT (20B)

2022

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



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



MARKETS TECHNOLOGY MONEY HEALTHCARE ENERGY MAPS MINING GREEN PARTNER WITH US

CREATOR PROGRAM

REPORTS CHANNELS



TECHNOLOGY

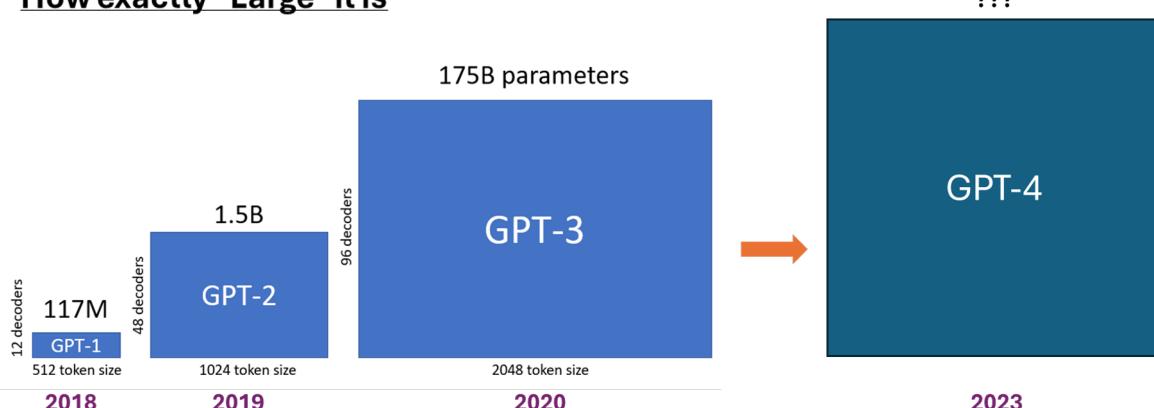
Visualizing the Training Costs of AI Models Over Time



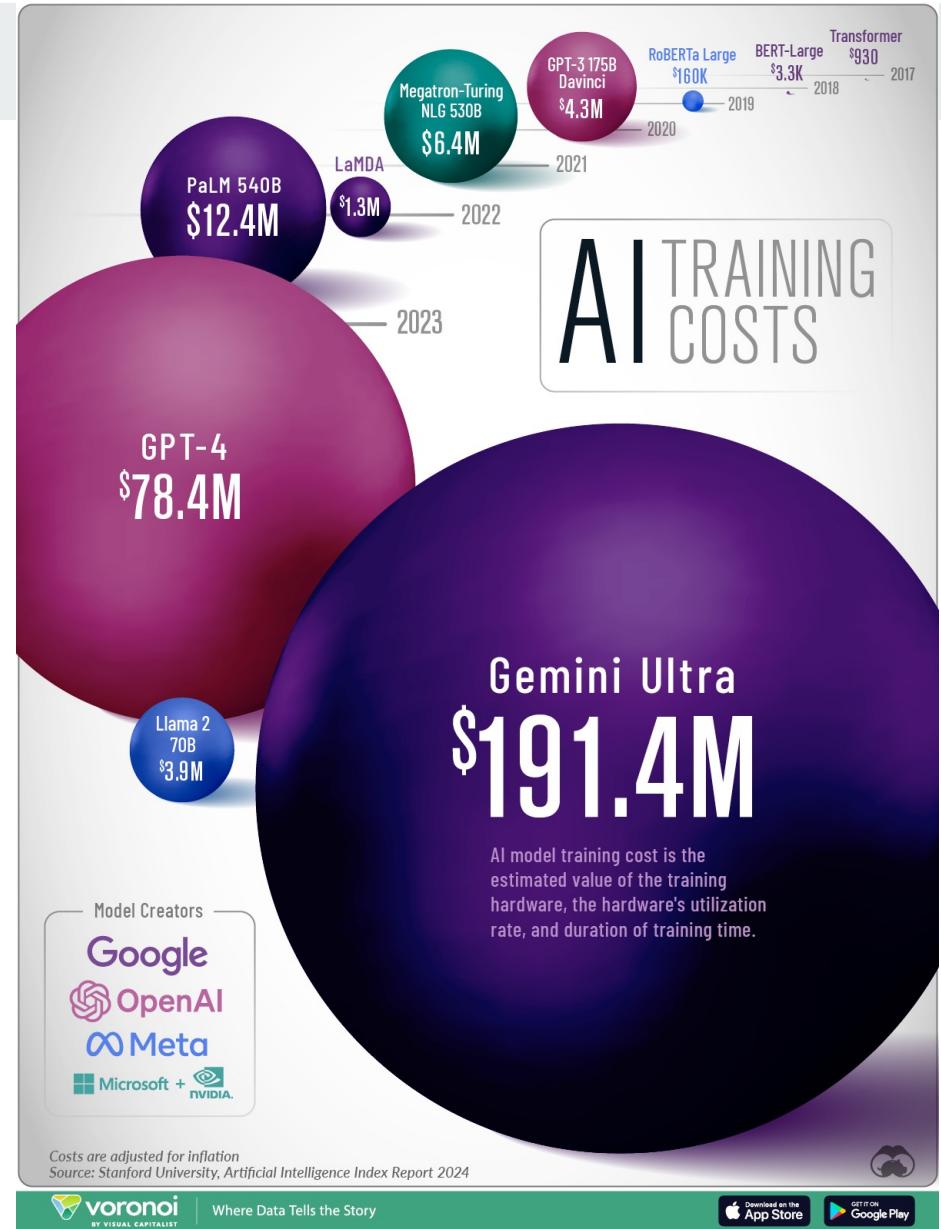
Published 3 months ago on June 4, 2024
By Dorothy Neufeld

- 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



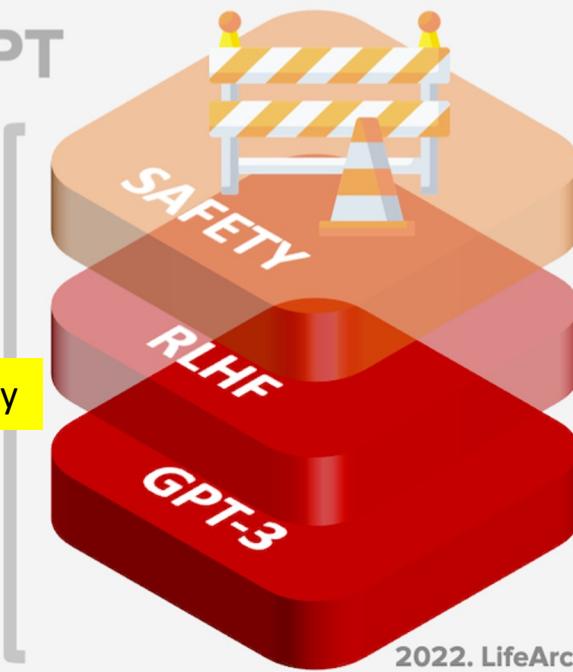
InstructGPT

Jan/2022



ChatGPT

Nov/2022



2022. LifeArchitect.ai

Step 1

Collect demonstration data and train a supervised policy.

LM

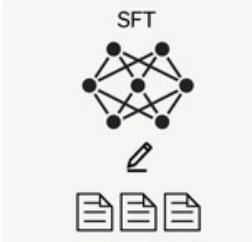
A prompt is sampled from our prompt dataset.



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)

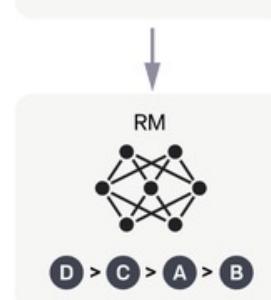
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

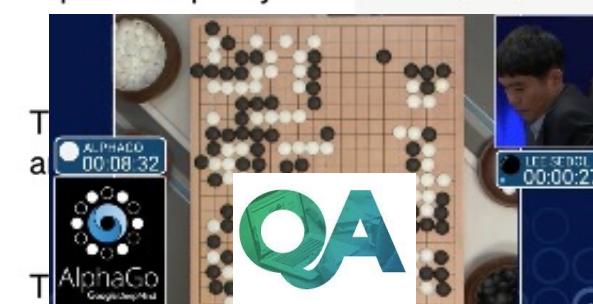
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

RL

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



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 on 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
18

<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 screenshot of the ChatGPT interface. It shows a sidebar with a "New chat" button, a "Previous 30 Days" section, and a "March" section listing "Sora: AI Text-to-Video Model", "Starlight Explorer: Magical Adventure", "WriteWise: Content Creation Revolution", and "AI Prompts Enhance Copywriting". The main area has a "ChatGPT" logo and the text "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 screenshot of the Google Gemini interface. It features a "Hello, Ayush." greeting and the question "How can I help you today?". Below this 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". At the bottom, there's an "Enter a prompt here" input field and a note about Gemini's privacy.
- Anthropic Claude**: A screenshot of the Anthropic Claude interface. It shows a "Hey" message, a "Claude by ANTHROPIC" logo, and the text "What can I help you with?". At the bottom is a "Message Claude..." input field with 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.

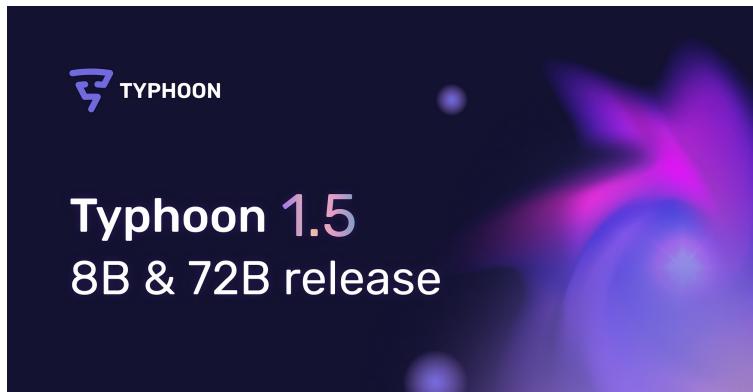
<https://openai.com/index/hello-gpt-4o/>

Commercial Large Language Model (LLM)

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

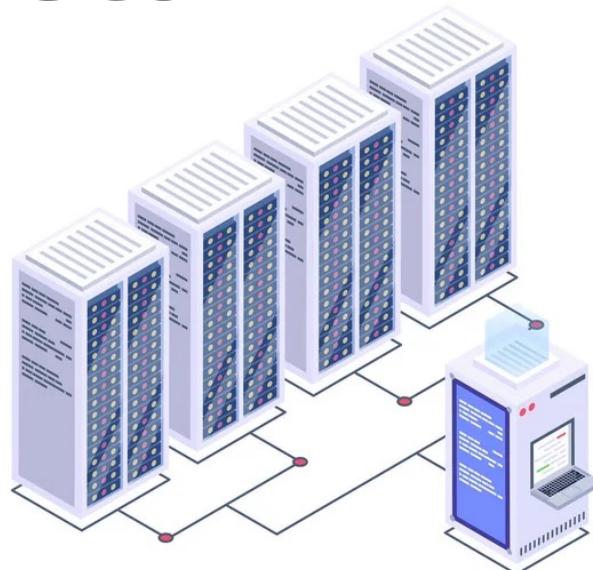
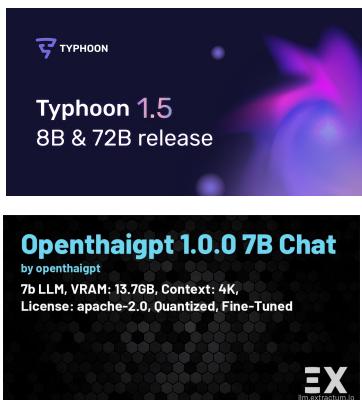


Thai LLMs: OpenThaiGPT, Typhoon, etc.



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

LLaMA
by  Meta



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

พูดคุยกับ "กนอย" ขออภัยเดี๋ยวเราจัดสุ่มภาษา AI ให้เป็นสากล
ด้วยพลังของ Chochae GPT

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

04:21

99+ หน่อย

ช่วยวิเคราะห์รูปคดี ตำรวจจับ 'ลุง เปี้ยก' จัดหลักฐานให้ 'ฝ่าเมีย' เตรียม ส่งข้อการอ่าน 'นิยายสอนสวน' ลั่งฟื้อง คดี มี 'แผนประทุยกรรม' ประกอบ เห็นมื่นเปี๊ะ!
ใช้เก้าอี้ตีจนตาย แต่คนร้ายตัวจริงกลับ ไม่ใช่ ใจร้ายจัดๆ ให้ 'ลุงเปี้ยก' รับสารภาพ 'ข้อหาฆ่า' แทน 'แก้กลุก ตำรวจ'?

10:02

คันพับนทความดังต่อไปนี้

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

จากการนี้เข้าใจว่าที่ทำการจับกุมตัว "ลุง เปี้ยก" และจัดจากให้รับสารภาพว่า เป็นผู้กระทำการข้อ控告น่อง โดยใช้เก้าอี้ ตีจนเดือด แต่จากการเรียนสวน สอนสวนในภายหลังกลับพบว่า คนร้าย ตัวจริงไม่ใช่ "ลุงเปี้ยก" แต่เป็นแก็งลูก ตำรวจ ซึ่งมีการวางแผนประทุยกรรม ให้ "ลุงเปี้ยก" รับสารภาพแทน

ในการนี้ "ลุงเปี้ยก" อาจมีสิทธิ์ฟ้อง ร้องดำเนินคดีกับเจ้าหน้าที่ตำรวจ ที่เกี่ยวข้องในความผิดฐาน

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

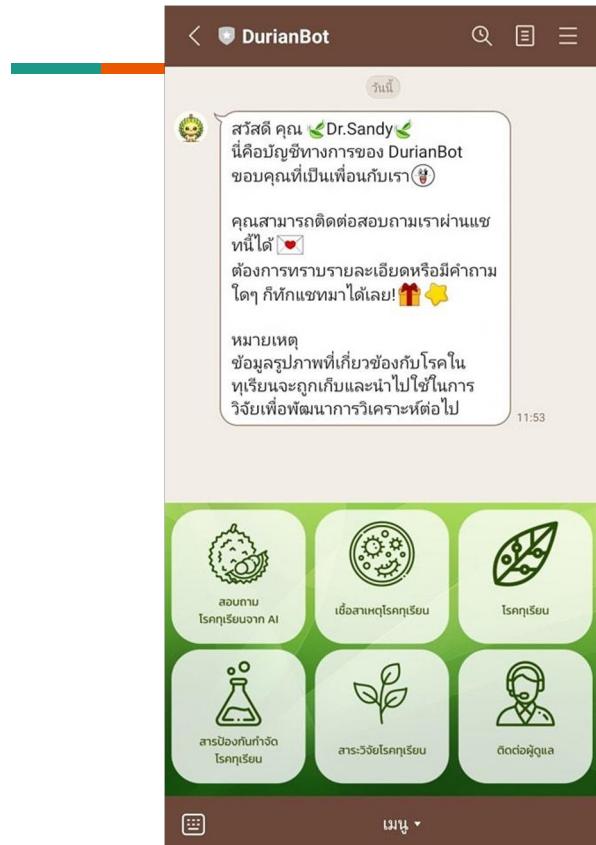
...

Use Cases in Agriculture



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

Use Cases in Agriculture



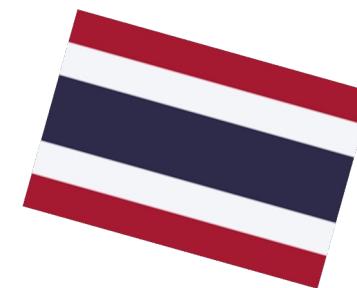
Reference: <https://www.durioinnovation.com/>, DurianBot, ไลน์บอทโรคข้าว

Practical chatbot with RAG

LLM กับภาพหลอน (hallucination)



 OpenAI
GPT-4



ChatGPT



ภาพหลอน (hallucination) คือการที่ LLM ตอบแบบให้ข้อมูลที่เหมือนจะจริง (มั่ว) จึงต้องระวังมาก ๆ โดยเฉพาะอย่างยิ่งงานด้านกฎหมาย การแพทย์ เป็นต้น ดังนั้น ก่อนใช้ LLM จึงควรต้องตรวจสอบก่อนใช้เสมอ

พีรพล เวทีกุล คือใคร



พีรพล เวทีกุล เป็นนักข่าวที่มีชื่อเสียงในประเทศไทย โดยเฉพาะในการสื่อสารมวลชน เขาเป็นที่รู้จักในฐานะผู้ประกาศข่าวทางช่อง 3 และเป็นพิธีกรรายการที่มีความนิยมในหลายรายการโทรทัศน์ เช่น รายการข่าวพิเศษ ต่าง ๆ และรายการสารคดีต่าง ๆ ที่ออกอากาศในช่อง 3 ของประเทศไทย



ณรงค์เดช กีรติพرانนท์ คือใคร



ณรงค์เดช กีรติพرانนท์ เป็นนักการเมืองและนักธุรกิจชาวไทยที่มีชื่อเสียง โดยเป็นสมาชิกพรรคการเมืองชั้นนำ ในประเทศไทย นอกจากนี้เขายังเป็นผู้บริหารในธุรกิจต่าง ๆ



Retrieval Augmented Generation (RAG)



Assistants with RAG in ChatGPT

P Personal / Default project

Playground Dashboard Docs API reference 

DASHBOARD Assistants + Create

Assistants

Today, Sep 5

LawBot 3:14 PM
asst_nntNyob1cbMmEpvcnbWdmG2a

ASSISTANT
asst_nntNyob1cbMmEpvcnbWdmG2a Playground ↗

Name
LawBot
asst_nntNyob1cbMmEpvcnbWdmG2a

Instructions
สมมติว่าเป็นนักกฎหมาย กำลังตอบค่ำถามกฎหมาย

Model
gpt-4o

TOOLS

File search ⓘ + Files
Vector store for LawBot
vs_QQS3iMZ3rtCgPDHwFwwz8DVi 16 KB

Code interpreter ⓘ + Files
Clone Updated 9/5, 3:24 PM

Forum Help

Adding documents to search engine database

Screenshot of the LangChain Playground interface showing the process of adding documents to a vector store for a LawBot.

The interface includes:

- Left sidebar:** Personal, Default project, PLAYGROUND, Chat, Assistants (selected), TTS, Completions, Forum, Help.
- Top navigation:** Playground, Dashboard, Docs, API reference, Settings, Profile.
- Central area:** Playground, LawBot, THREAD, Vector store for LawBot modal.
- Vector store for LawBot modal:**
 - Name:** LawBot
 - Instructions:** สมมติว่าเป็นนักกฎหมาย กำลังดูคดีกฎหมาย
 - Model:** gpt-4o
 - Tools:** File search, Code interpreter.
 - Vector store details:** ID: vs_QQS3iMZ3rtCgPDHwFwwz8DVi, Expiration policy: Never.
 - File upload history:** Peerapon_ThaiCV.pdf (9/5/2024, 3:57 PM) and HeritageLaw 2.pdf (9/5/2024, 3:21 PM).
 - Buttons:** + Add, Detach from assistant.
- Bottom right overlay:**
 1. ประวัติส่วนตัว

ชื่อและนามสกุลภาษาไทย : รองศาสตราจารย์ ดร. พีรพล เวทีกุล
ชื่อและนามสกุลภาษาอังกฤษ : Associate Professor Dr. Peerapon Vateekul


- Playground messages can be viewed by anyone at your organization using the API.

More reliable answer with RAG

Personal / Default project

Playground Dashboard Docs API reference ⚙️ 

PLAYGROUND Chat Assistants TTS Completions

Playground

LawBot

THREAD thread_8CPr6e303DjVgRfCsDjTo4WR 6373 tokens

Name LawBot asst_nntNyob1cbMmEpvcnbWdmG2a

Instructions สมมติว่าเป็นนักกฎหมาย กำลังตอบคำถามกฎหมาย

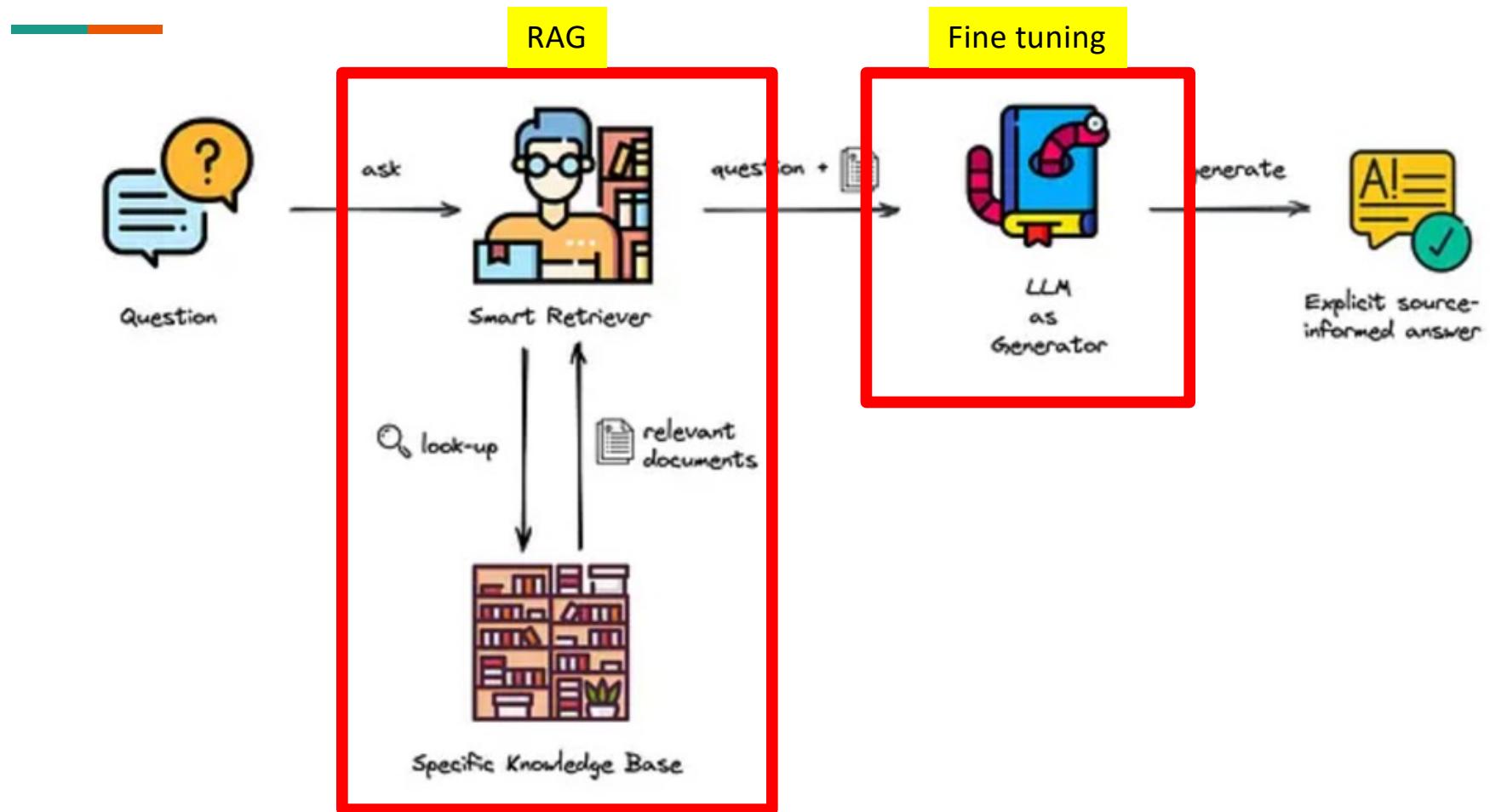
Model gpt-4o

TOOLS File search (Vector store for LawBot vs_QQS3iMZ3rtCgPDHwFwwz8DVi 16 KB) + Files Code interpreter (+ Files) + Run ⌘ ↵

Enter your message...  

Playground messages can be viewed by anyone at your organization using the API.

RAG vs Fine tuning



Q&A

Peerapon.v@chula.ac.th