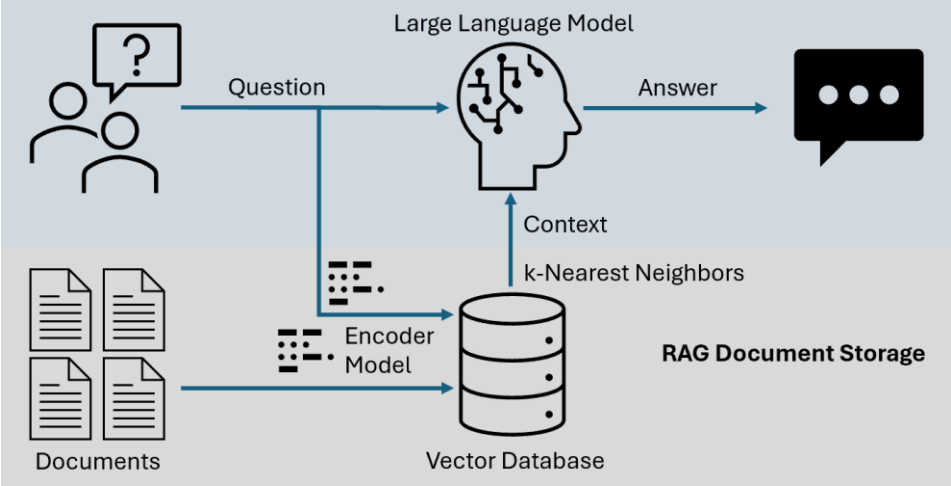
簡介大型語言模型

https://llm-chronicles.com/

https://medium.com/@henryhengluo/intro-of-ai-agent-ai-agent-projects-summary-52f4a364ab86

11/10/2024
Pslams Cafe



LLM + RAG 是大腦

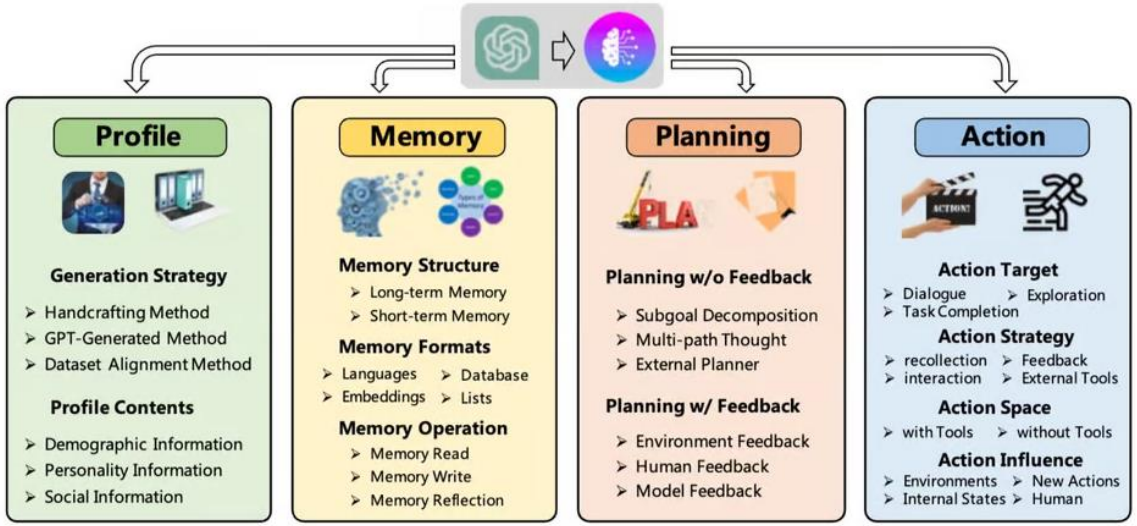
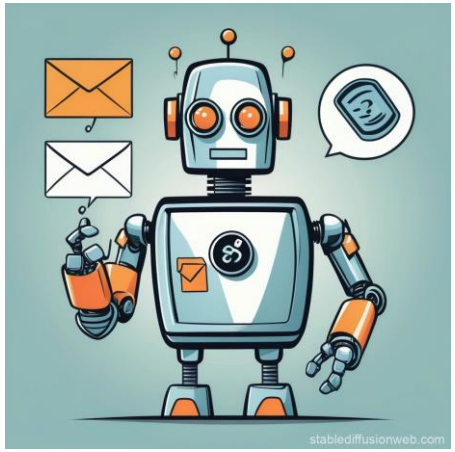


Figure 2: A unified framework for the architecture design of LLM-based autonomous AI agent.

AI agent 是機器助理
去執行各種任務



Outline:

- ChatGPT 是什麼？
- Transformer是什麼？
- Encoder/Decoder是什麼？
- AI 發展史
- LLM + **RAG** 的應用
- AI代理（AI Agent）的應用

ChatGPT是什麼？

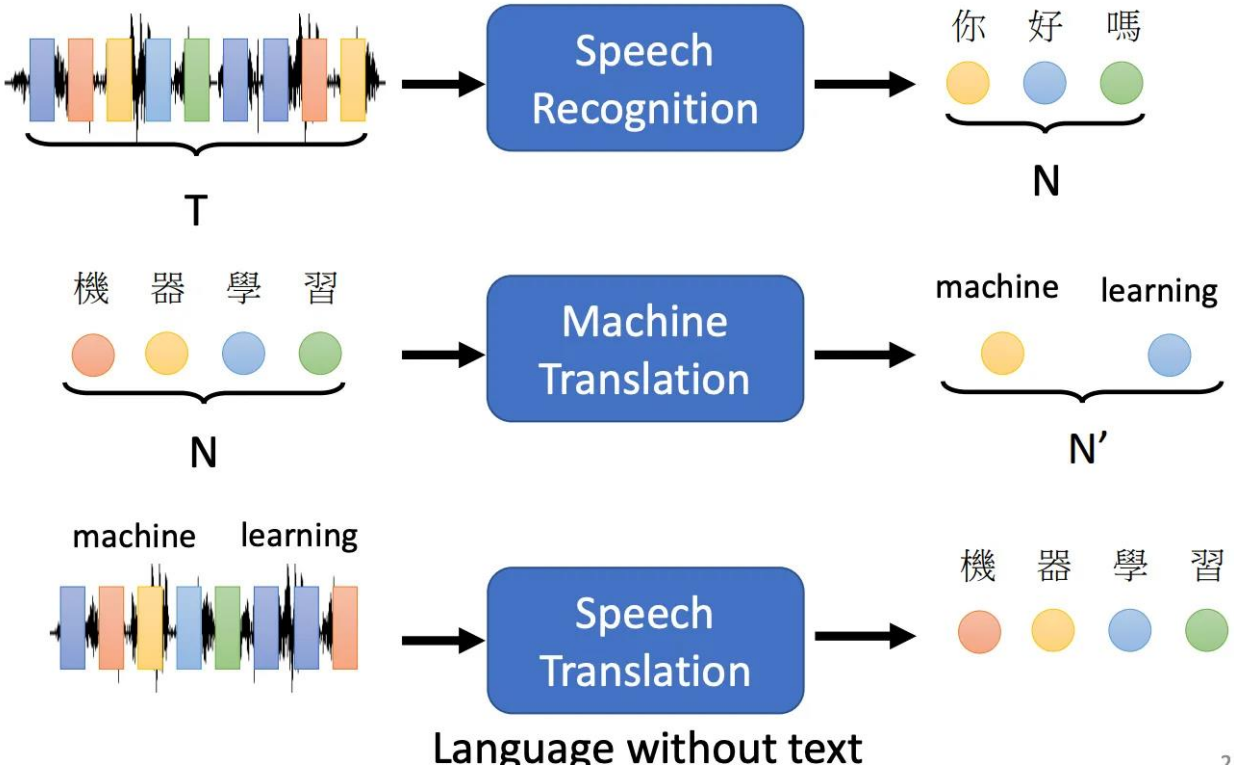
- ChatGPT是一款聊天機器人，它會生成類似人類會寫出來的文字。ChatGPT可以自然地回答眾多問題、精通許多學科，就像個私人導師一樣。
- ChatGPT是「生成式預訓練轉換器」（Generative Pre-Trained Transformer）技術的最新發展。
- 它採用深度學習（deep learning），根據從網路上獲取的大量文本樣本進行訓練。

Transformer 是什麼?

Sequence-to-sequence (Seq2seq)

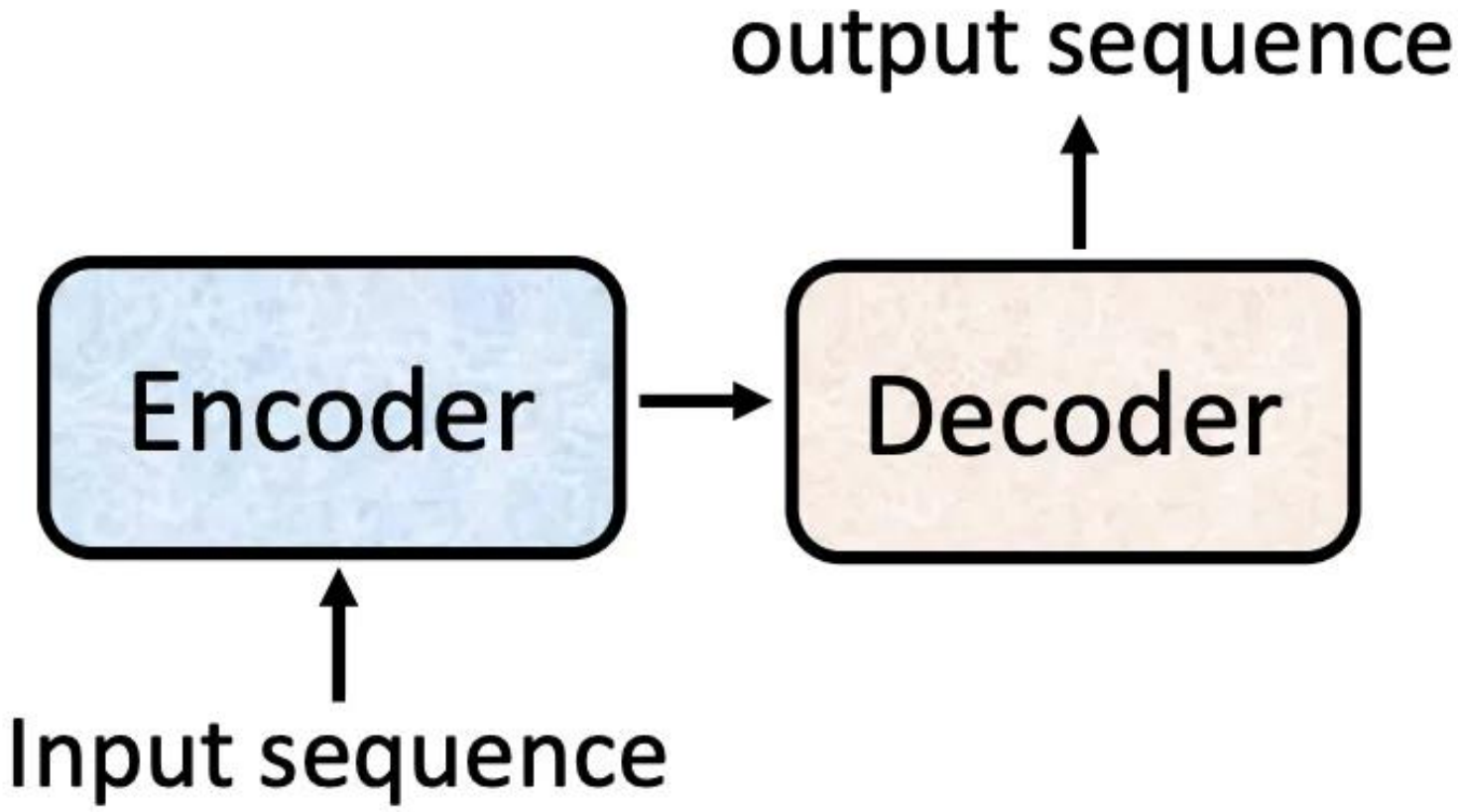
Input a sequence, output a sequence

The output length is determined by model.



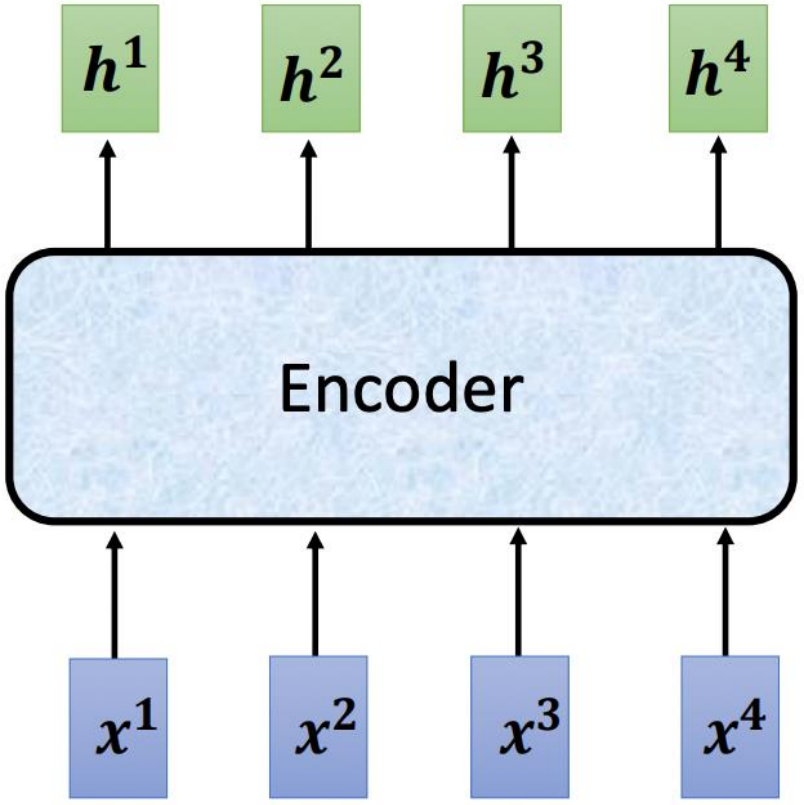
Transformer即為一個Sequence to sequence(Seq2seq)的model, 由**機器自己決定output的長度**!

Transformer 是什麼?

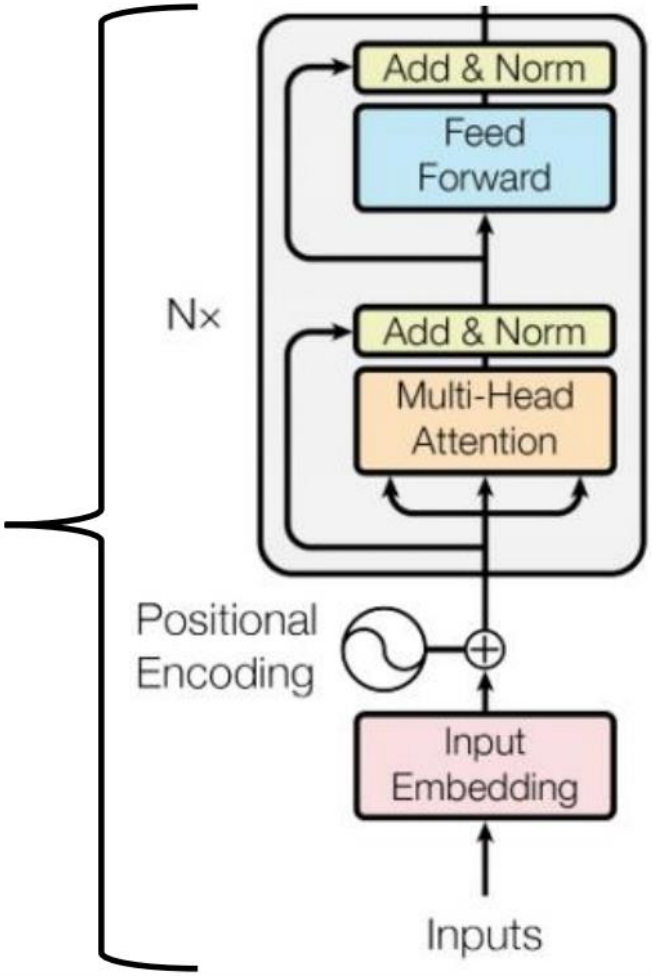


Encoder

You can use **RNN** or **CNN**.

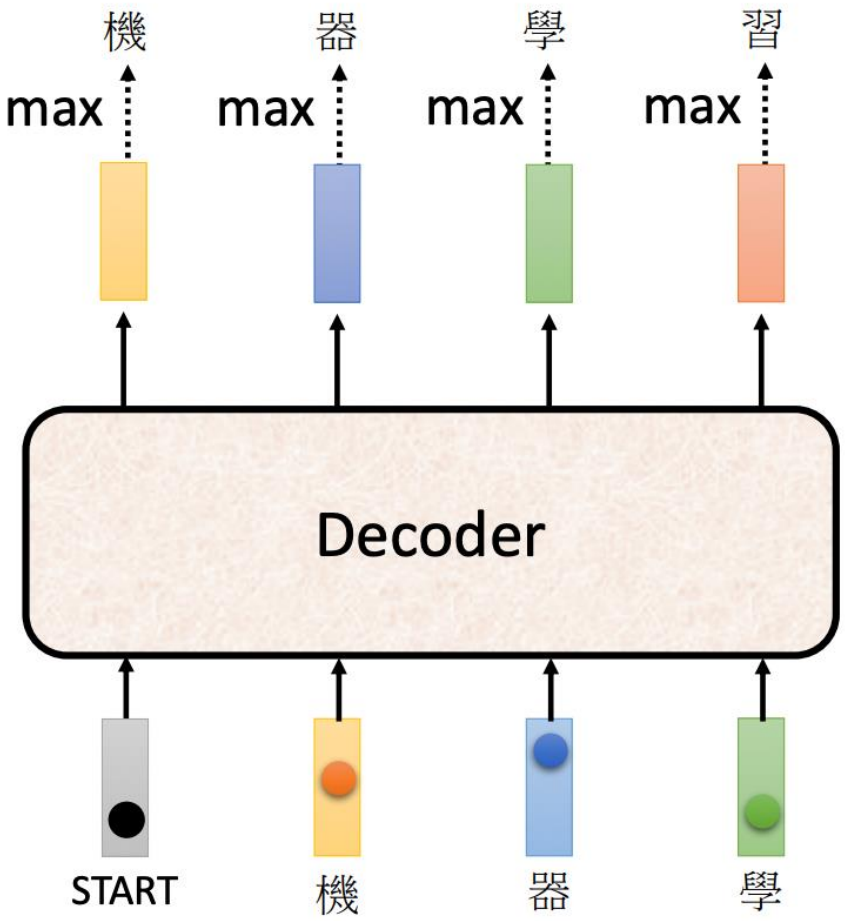


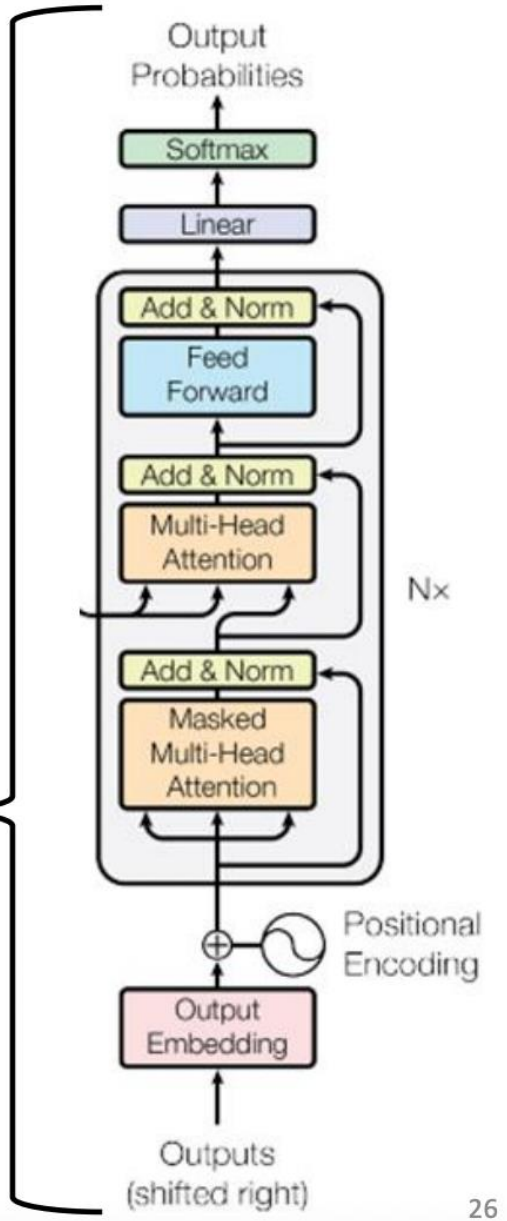
Transformer's Encoder



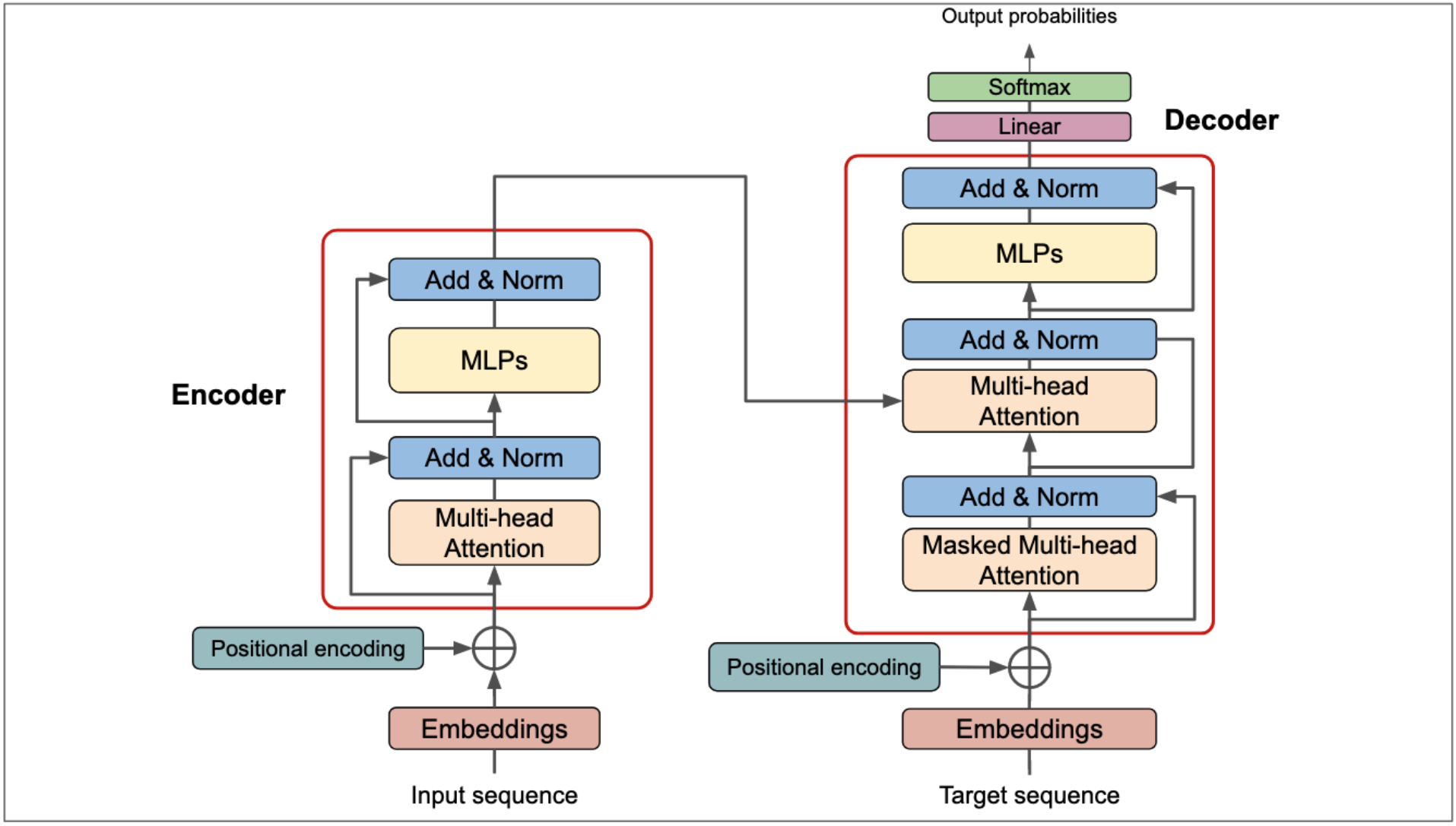
Decoder

ignore the input from the encoder here ☺





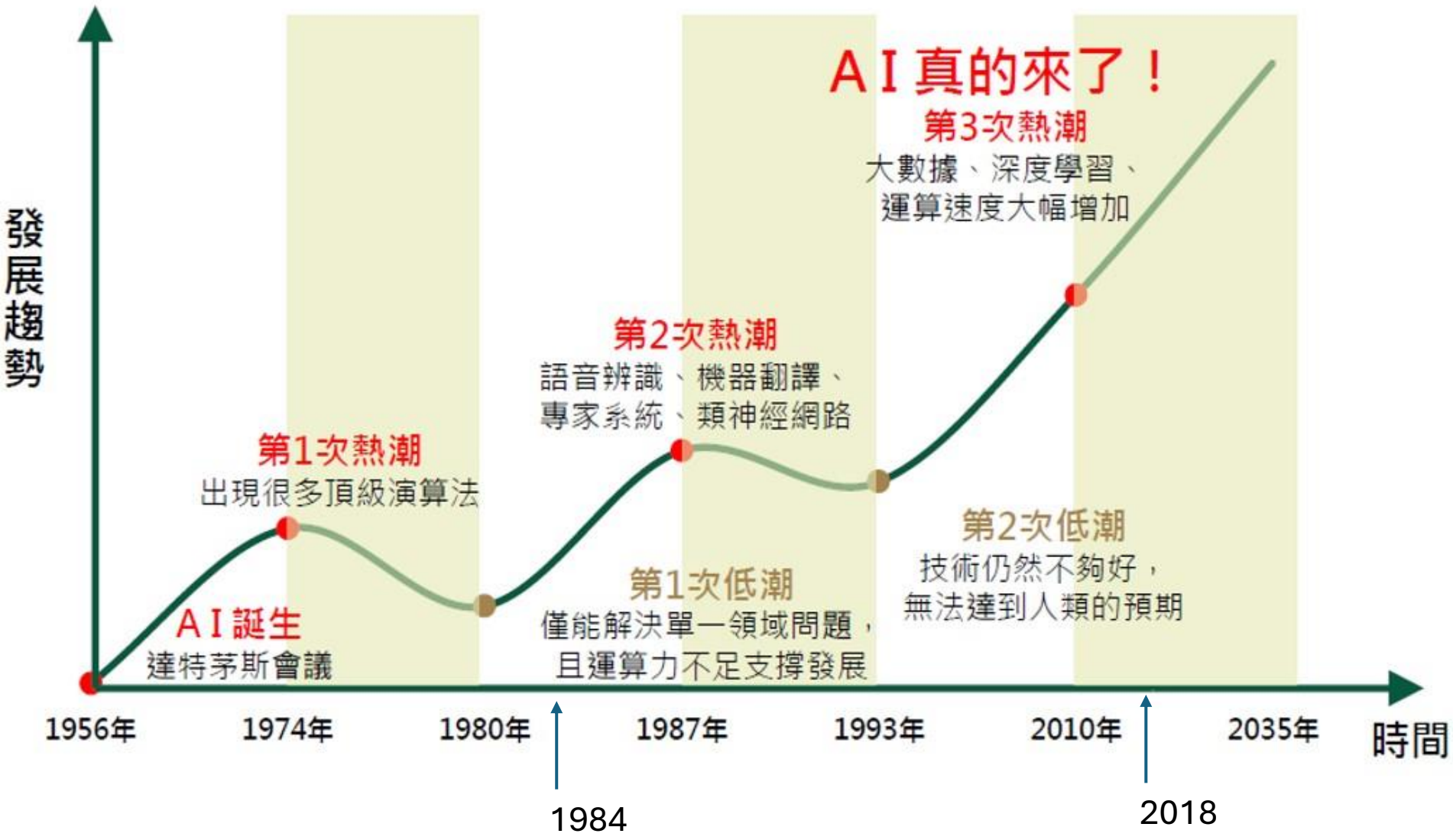
Transformer 是什麼?



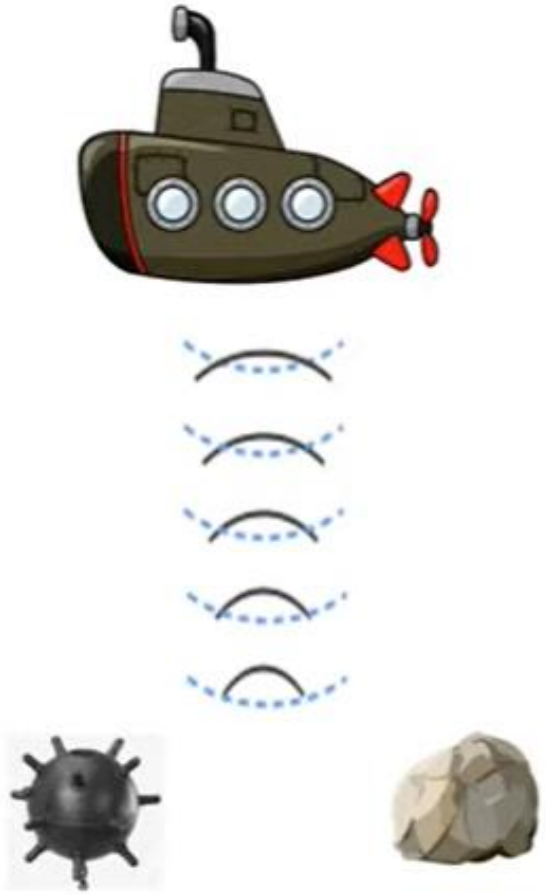
Outline:

- ChatGPT 是什麼？
- Transformer是什麼？
- Encoder/Decoder是什麼？
- AI 發展史
- LLM + **RAG** 的應用
- AI代理（AI Agent）的應用

AI發展史


發展趨勢
AI真的來了！
第3次熱潮
大數據、深度學習、
運算速度大幅增加
第2次熱潮
語音辨識、機器翻譯、
專家系統、類神經網路
第1次熱潮
出現很多頂級演算法
第2次低潮
技術仍然不夠好，
無法達到人類的預期
第1次低潮
僅能解決單一領域問題，
且運算力不足支撐發展
AI誕生
達特茅斯會議
1956年
1974年
1980年
1984
1987年
1993年
2010年
2018
2035年
時間

1984 的 AI



SONAR
Rock vs Mine Prediction
With Python

Neural Networks, Vol. 1, pp. 75–89. 1988
Printed in the USA.

0893-6080/88 $3.00 + .00

ORIGINAL CONTRIBUTION

Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets

R. PAUL GORMAN

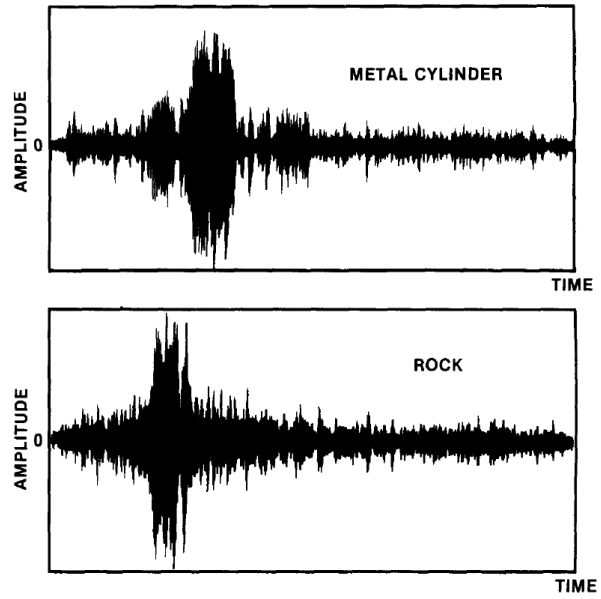
Allied-Signal Aerospace Technology Center

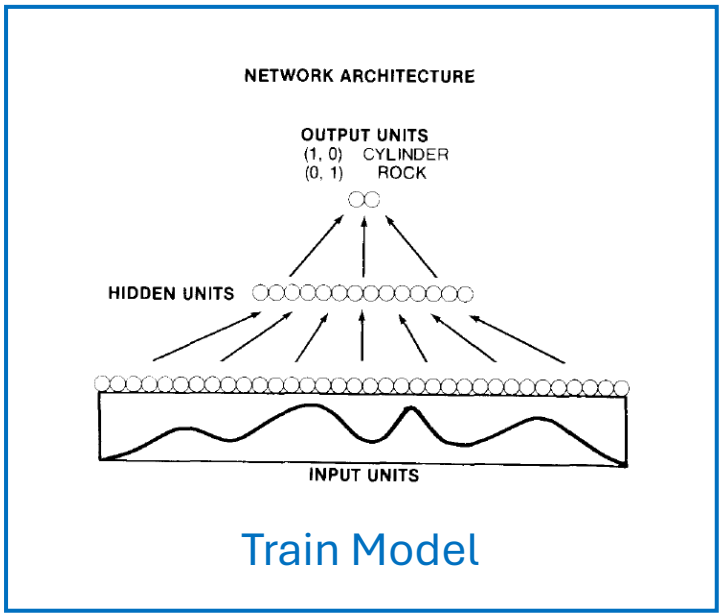
TERRENCE J. SEJNOWSKI

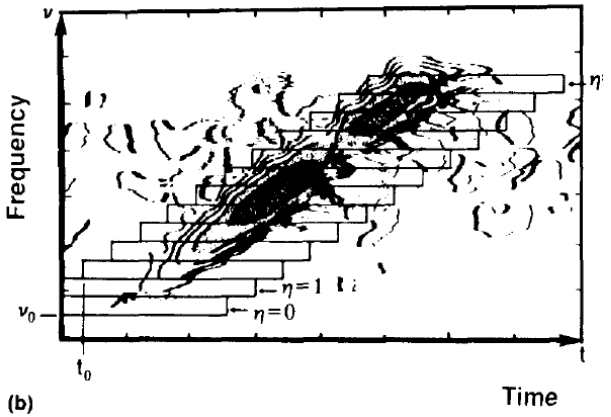
Johns Hopkins University

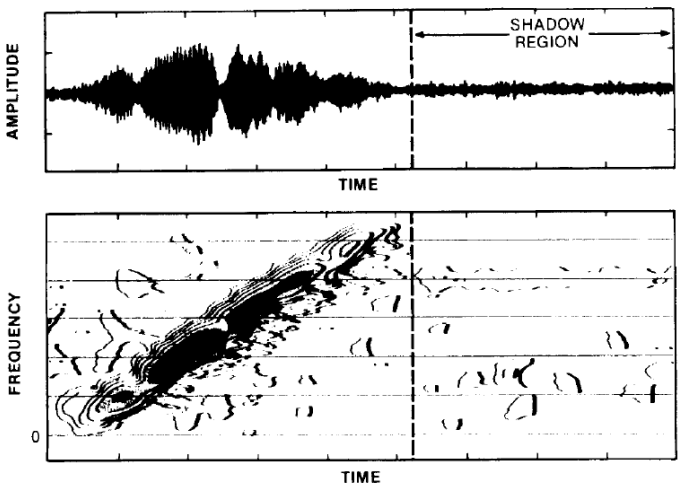
(*Received and accepted* 30 *October* 1987)

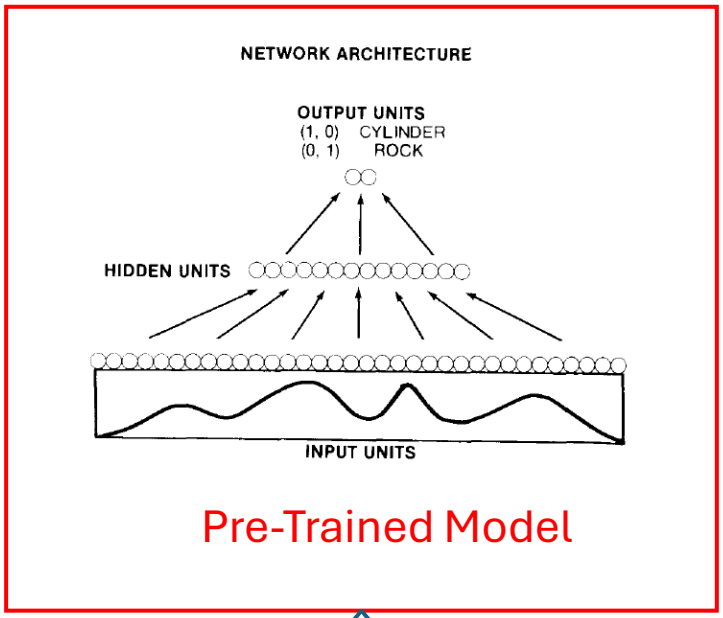
Abstract—*A neural network learning procedure has been applied to the classification of sonar returns from two undersea targets, a metal cylinder and a similarly shaped rock. Networks with an intermediate layer of hidden processing units achieved a classification accuracy as high as 100% on a training set of 104 returns. These networks correctly classified up to 90.4% of 104 test returns not contained in the training set. This performance was better than that of a nearest neighbor classifier, which was 82.7%, and was close to that of an optimal Bayes classifier. Specific signal features extracted by hidden units in a trained network were identified and related to coding schemes*


METAL CYLINDER
AMPLITUDE
0
TIME
ROCK
AMPLITUDE
0
TIME

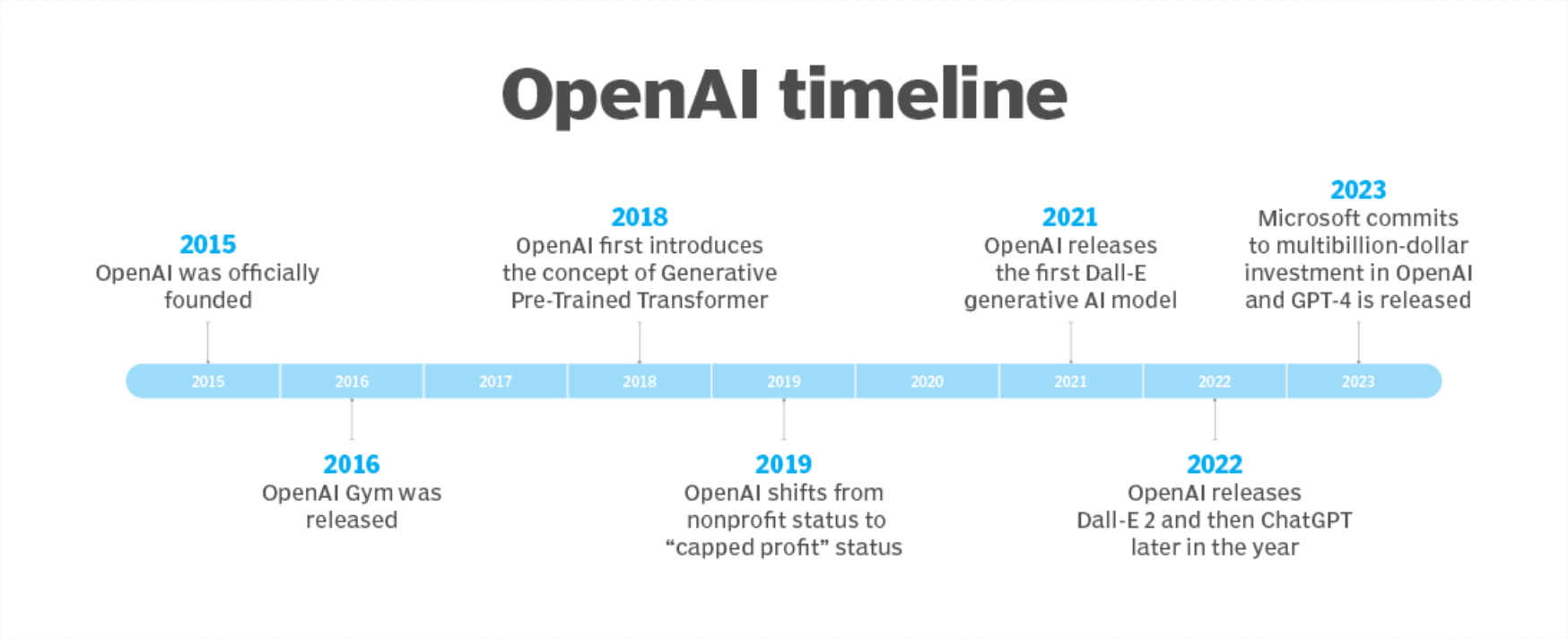

NETWORK ARCHITECTURE
OUTPUT UNITS
(1, 0) CYLINDER
(0, 1) ROCK
HIDDEN UNITS
INPUT UNITS
Train Model


Frequency
Time
(b)


AMPLITUDE
SHADOW
REGION
TIME
FREQUENCY
0
TIME


NETWORK ARCHITECTURE
OUTPUT UNITS
(1, 0) CYLINDER
(0, 1) ROCK
HIDDEN UNITS
INPUT UNITS
Pre-Trained Model

2015 民間新創OpenAI



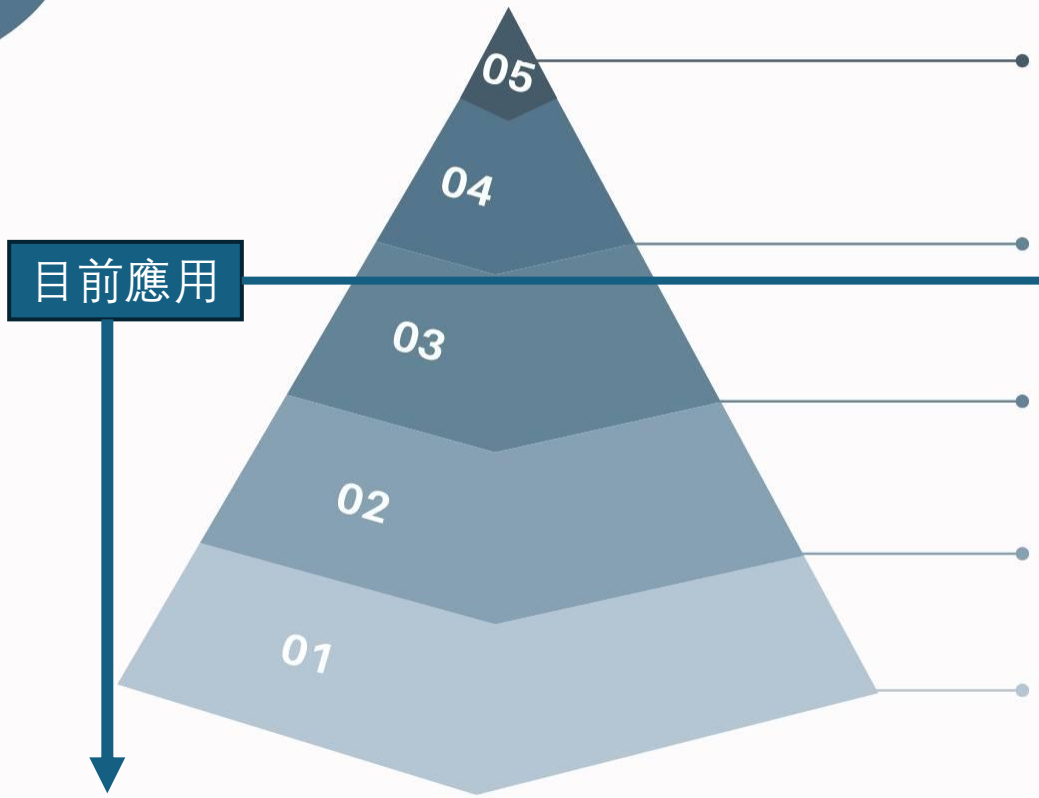
2017 政府介入 AI發展策略

全球各國政府 AI 發展策略

國家	說明
美國	2018年成立「AI 特別委員會」，向白宮提供 AI 研究發展的建議，並幫助政府、私人企業和獨立研究者建立合作夥伴關係，鞏固 AI 優勢地位
中國	2017年將AI列為國家發展戰略，將投注**1,500億美元**進行研發 **2030年**中國成為全球 AI 創新中心
日本	2018/6公布「日本未來投資策略2018」，重點培育 AI 人才 **2022年前**投入**100億日圓**打造10加「AI 醫院」
歐盟	2018/4加碼對 AI 的投資，將於2018~2020年間提供**240億美元**
南韓	**2018年**加碼**2.2兆韓圜**投資 AI 研發，2022年前創設6個 AI 研究院，培育5000名 AI 專家， **2026年**將相關技術研發提升到已開發國家水準
加拿大	2017年成為全球首個發布 AI 全國戰略的國家，投注1.25億加幣

2024 AI應用

OpenAI 通用人工智慧五級標準



組織 ORGANIZATIONS

最高級別的AI能夠執行複雜的組織任務，並在各種任務上超越所有人類的能力。這一級別的AI可以完成超越人類極限的工作，如預測未來、與動物交流等，達到科幻中的智能水平。

創新者 INNOVATORS

AI能夠自主進行發明和創新。這意味著AI不僅能完成既定任務，還能提出新的解決方案，幫助人類開拓新的技術和思路。

代理者 AGENTS

代理者級別的AI，能為用戶採取行動，完成一系列任務的AI。不僅能理解和推理，還能實際執行具體操作，就像是你專屬的個人助理或秘書。

推理者 REASONERS

它們將能夠解決人類等級的問題。表示這類型的 AI 可以進行基本的推理和問題解決能力。

聊天機器人 CHATBOT

AI能夠使用自然語言進行對話，即 AI 能夠用人類語言，進行基本交流互動。

Outline:

- ChatGPT 是什麼？
- Transformer是什麼？
- Encoder/Decoder是什麼？
- AI 發展史
- LLM + RAG 的應用
- AI代理（AI Agent）的應用

什麼是RAG?

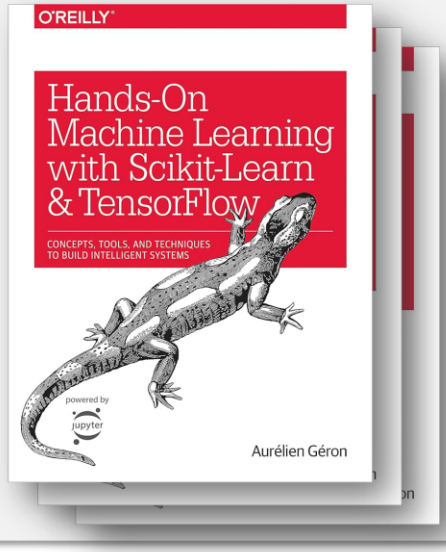
- RAG（Retrieval-Augmented Generation）由 Patrick Lewis 等人於 2020 年提出
- 是一種 AI 框架，旨在通過提供外部資料知識來提升 LLM（大型語言模型）的回答質量和準確性。

什麼是RAG?

Query

Query: 為什麼 ML 需要正規化？

Open Book：



Retrieval 找到相關章節

Query: 為什麼 ML 需要正規化？

語音搜尋
檢索答案

Context

Chapter 1
1 One-Hot Encoding
2. Dropout
3. Regularization
4. Missing Data
Chapter 2
1. CNN
2. RNN
3. LSTM
4. Transformer

Augmented-Generation 基於資料回答

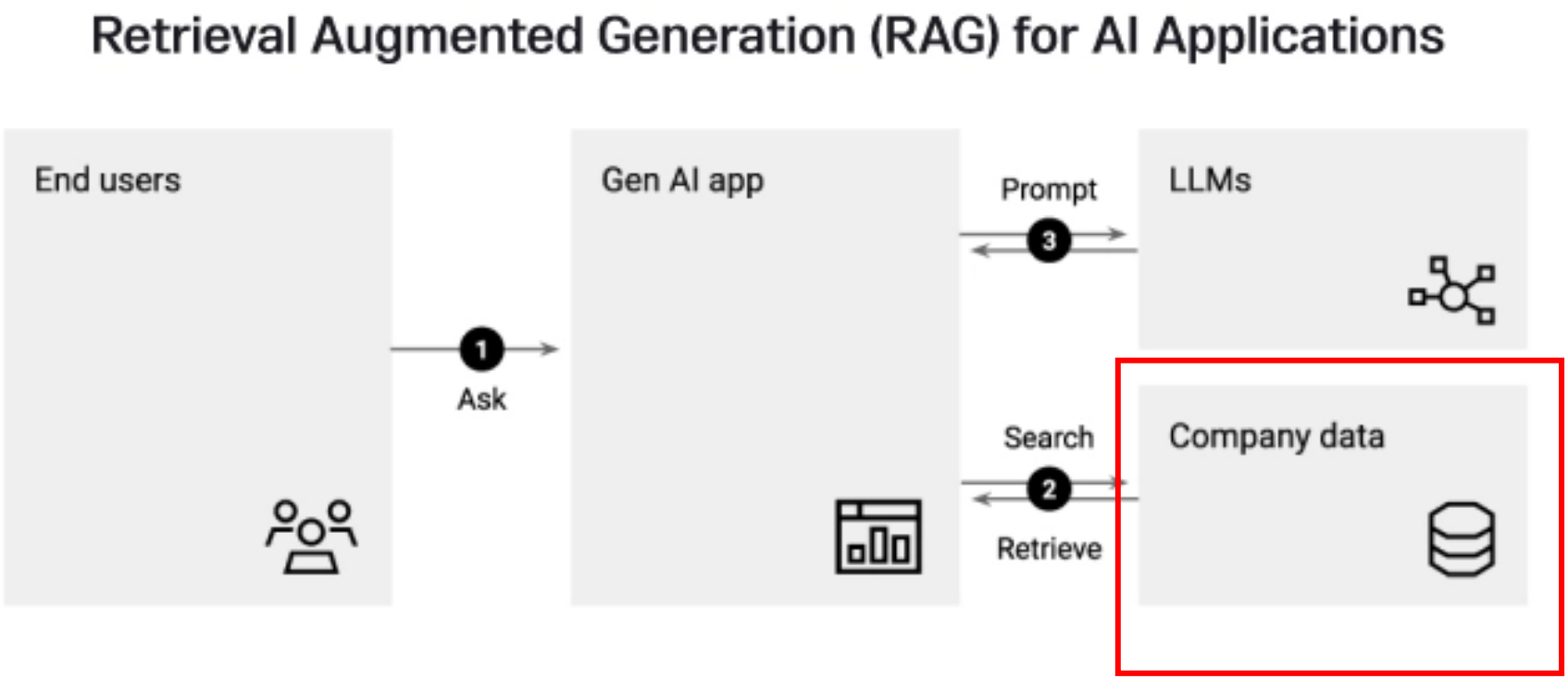
Query: 為什麼 ML 需要正規化？

+Retrieval Information:

1.3 Regularization：
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxx

Answer: 防止 Overfitting

使用生成式AI 與檢索增強生成（RAG）



大型語言模型

Models

RAG平台

RAG Frameworks

大型語言模型

開源 可以 Local 跑的模型

- 公開模型架構與權重，
  提供開發者下載、Fine-tune
- 需考量
  1. 模型大小 2. 語言 3. 電費成本 4. 可否商用

模型 Models

Llama 3
T5
Bloom
Mixtral

支援中文
中文 LLMs 排行

ChatGLM
Qwen
Breeze
TAIDE

平台 Models Hub

Hugging Face Hub
開源商用模型庫

Ollama Models
量化開源模型庫
支持 CPU 運算

閉源 只能 call API

- 沒有提供模型架構等細節
  開發者藉由官方 API 做模型調用
- 需考量
  1. API 價錢 2. 數據隱私 3. 可否商用

模型 Models

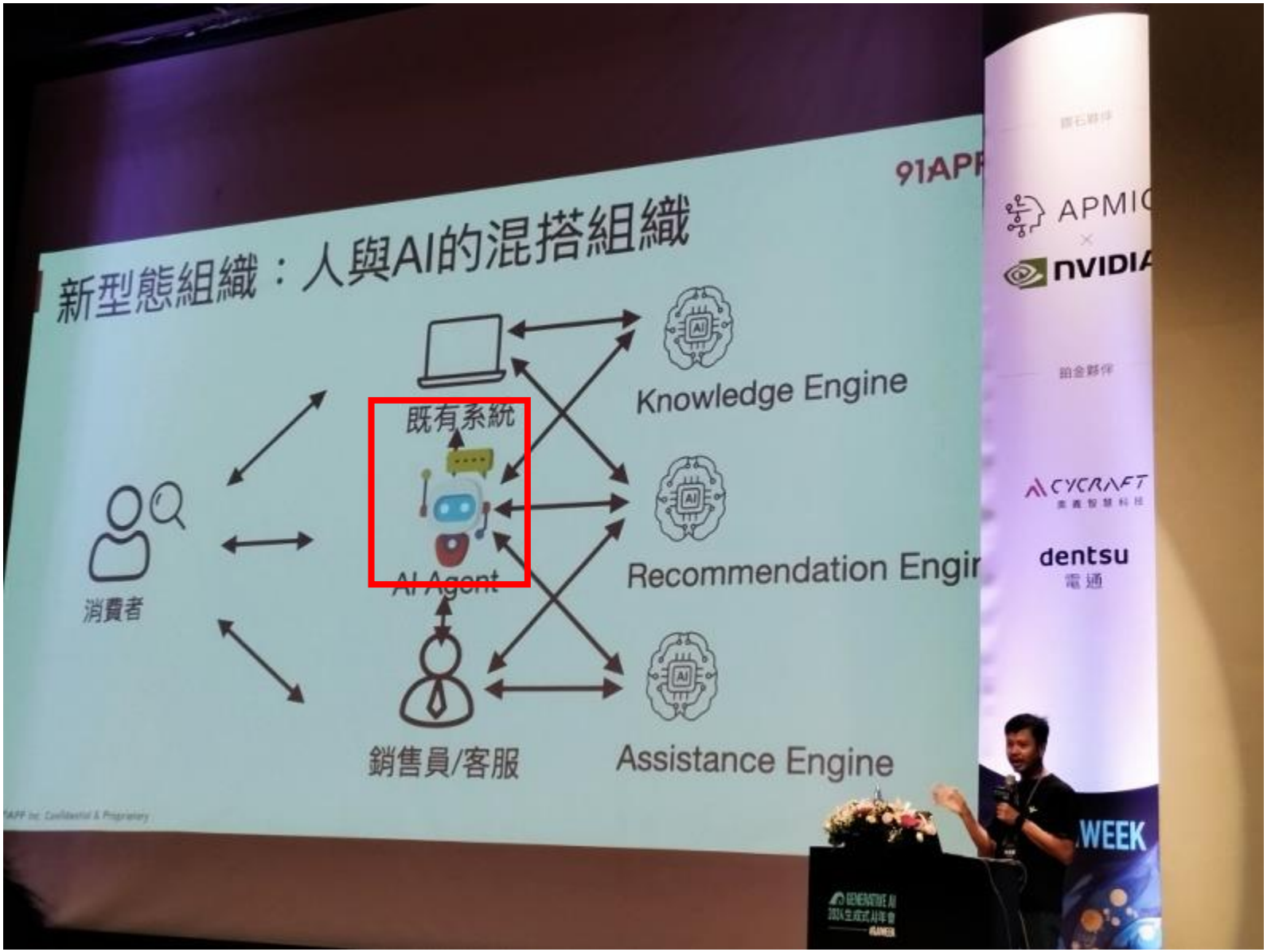
2024/04 資料，僅供參考

names	公司	上下文長度	價錢
gpt-4-tubo	OpenAI	32K	$10 /per
Gemini-1.5-pro	Google	1M	$7 /per
Claude 3	Anthropic	200K	$0.25 /per
Command R+	Cohere	128K	$3 /per

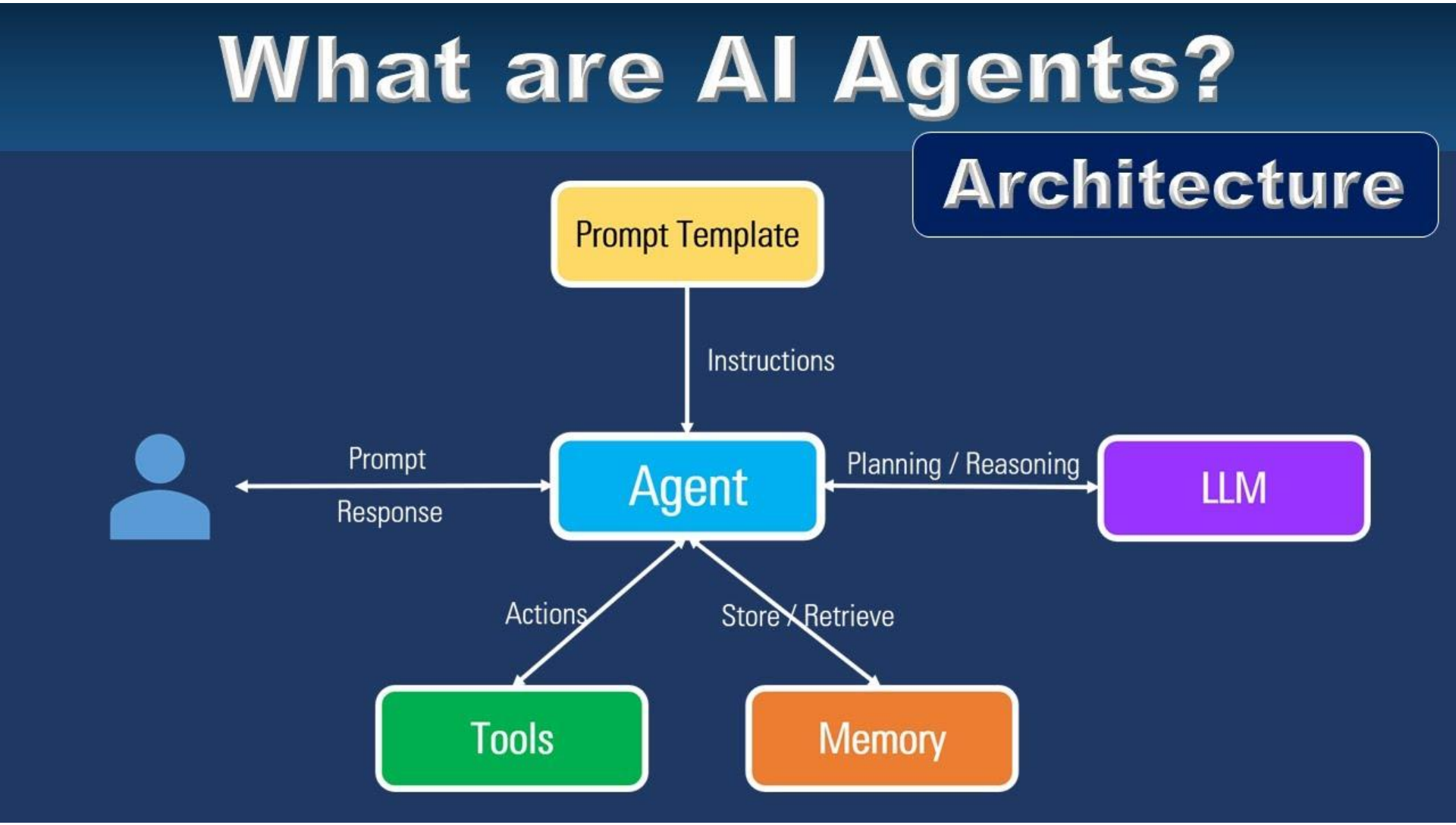
Outline:

- ChatGPT 是什麼？
- Transformer是什麼？
- Encoder/Decoder是什麼？
- AI 發展史
- LLM + **RAG** 的應用
- **AI代理（AI Agent）的應用**

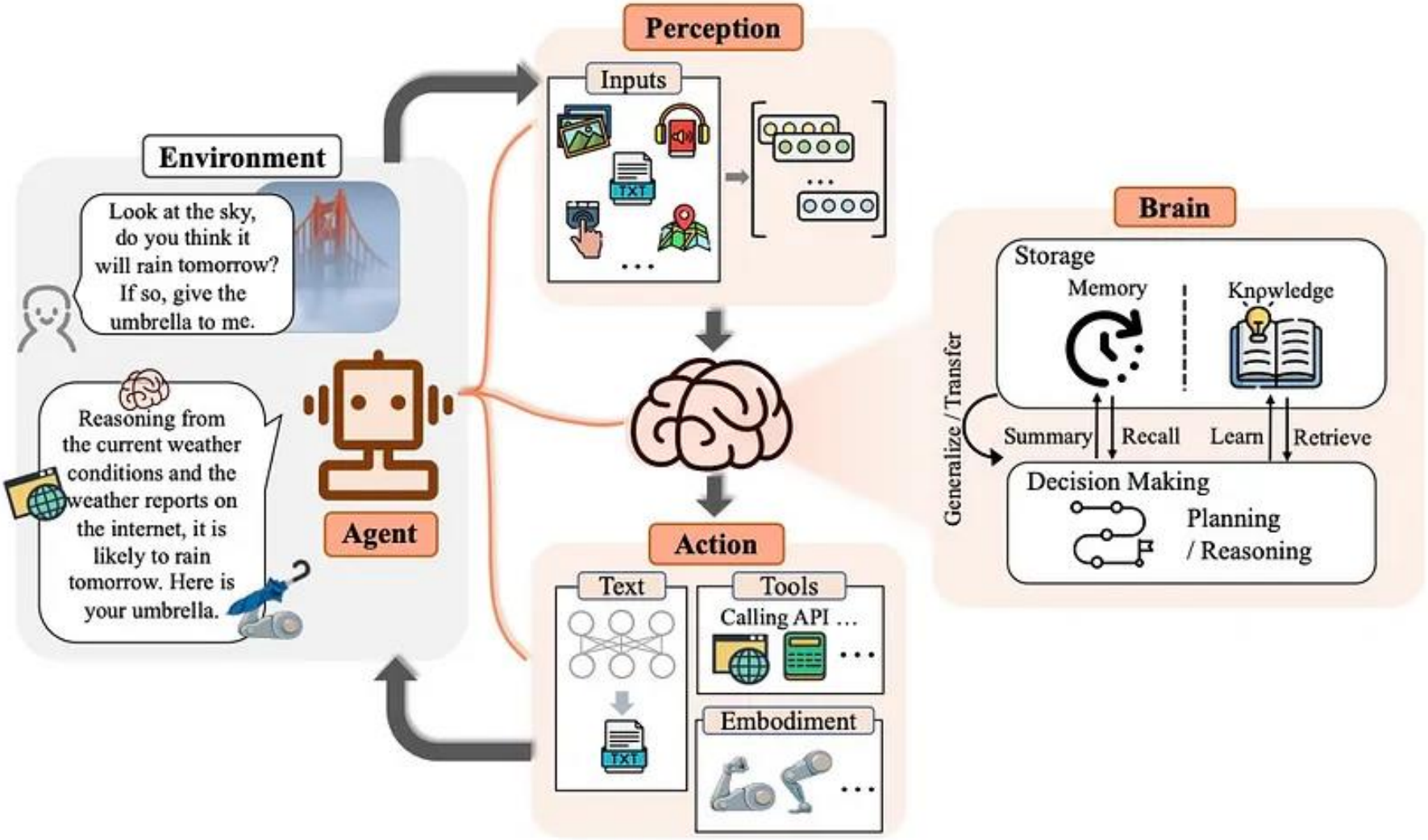
AI代理 （AI Agent）的應用



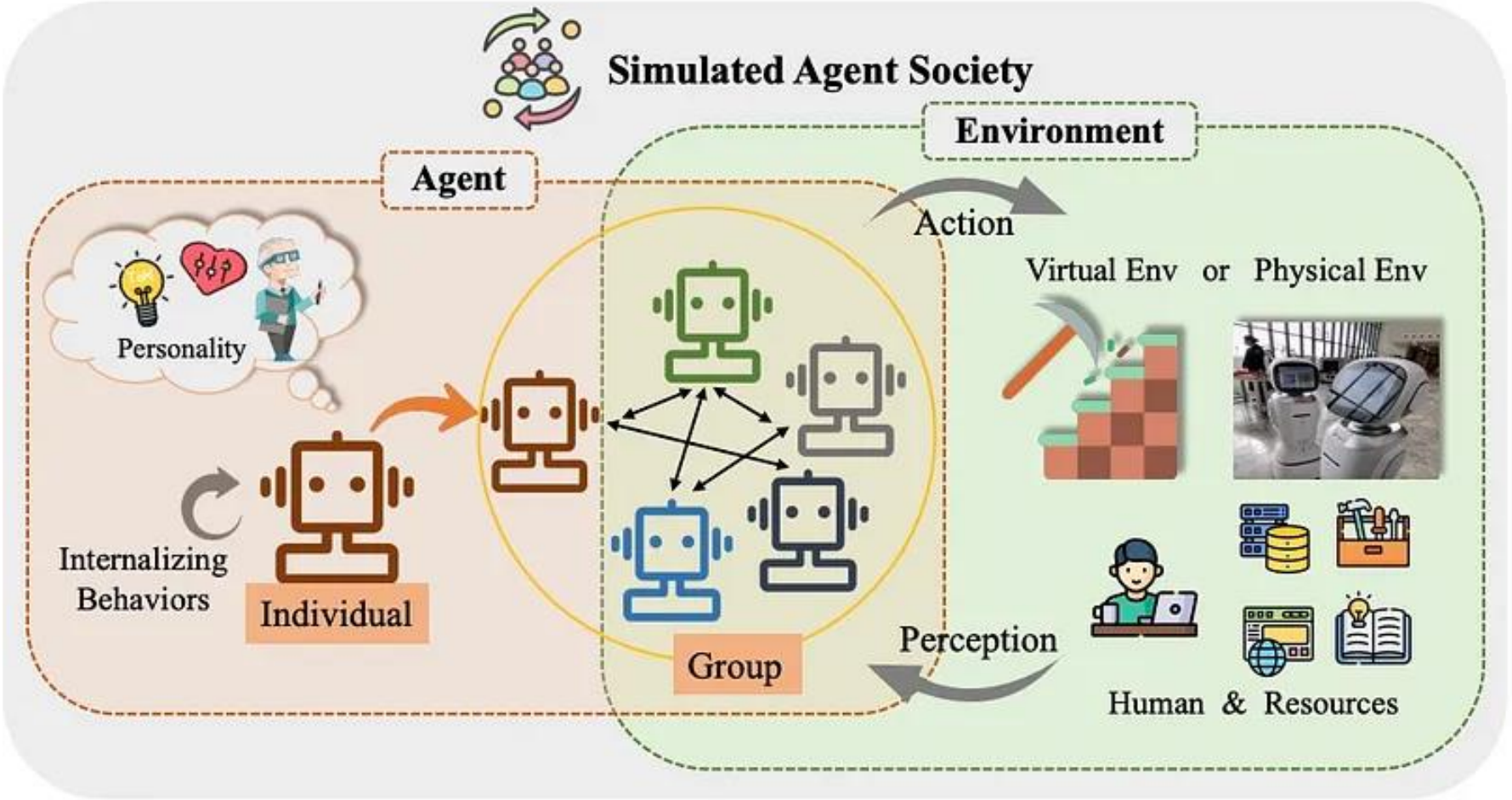
AI代理是什麼?


What are AI Agents?
Architecture
Prompt Template
Instructions
Prompt
Response
Agent
Planning / Reasoning
LLM
Actions
Store / Retrieve
Tools
Memory

AI代理是什麼?


Perception
Inputs
Environment
Look at the sky, do you think it will rain tomorrow? If so, give the umbrella to me.
Reasoning from the current weather conditions and the weather reports on the internet, it is likely to rain tomorrow. Here is your umbrella.
Agent
Action
Text
Tools
Calling API ...
Embodiment
Brain
Storage
Memory
Knowledge
Summary
Recall
Learn
Retrieve
Generalize / Transfer
Decision Making
Planning
/ Reasoning

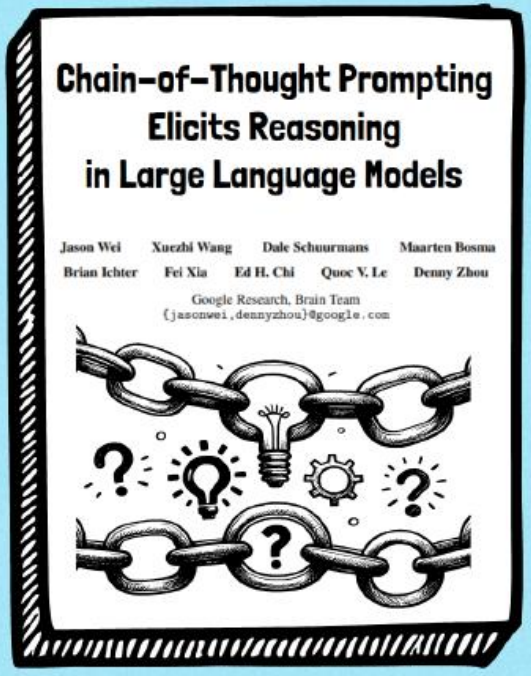
AI代理是什麼?


Simulated Agent Society
Environment
Agent
Action
Virtual Env or Physical Env
Personality
Internalizing Behaviors
Individual
Group
Perception
Human & Resources

AI Agent 4 大 papers

- Chain-of-Thought
- Zero-shot reasoner
- Program Aided Language Model
- ReAct : Reasoning and Acting

GOOGLE, 2022



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Think in INTERMEDIATE STEPS

INPUT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Example answer with step-by-step reasoning provided in the context

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

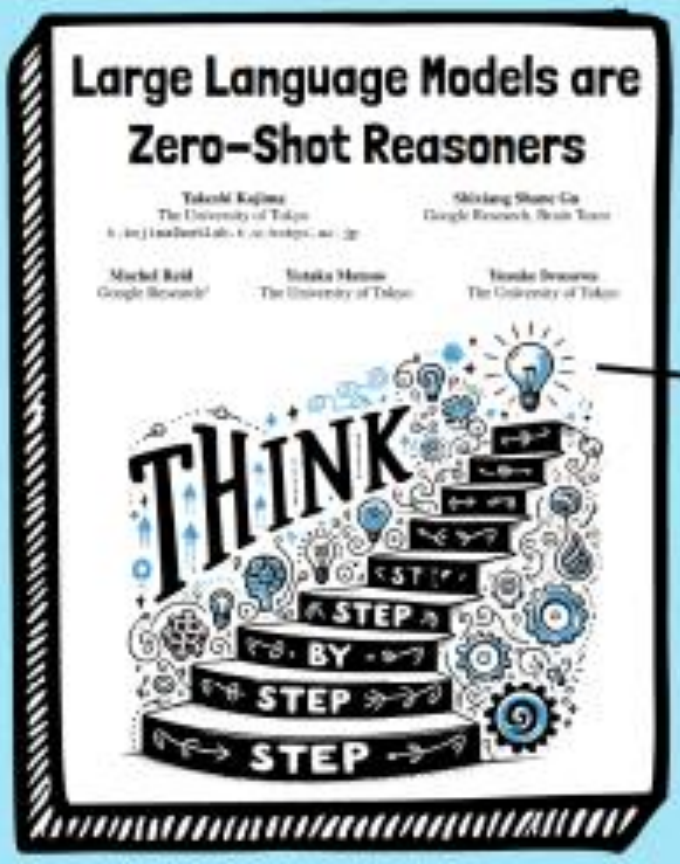
New question

OUTPUT

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

LLM reproduces step-by-step reasoing, resulting in better performance

UNIVERSITY OF TOKYO
GOOGLE, 2022



ZERO-SHOT CoT

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

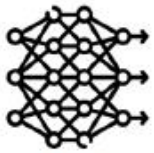
A: Let's think step by step.

There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

CMU, 2022

PAL: Program-aided Language Models

Luyu Gao[*1] Aman Madaan[*1] Shuyan Zhou[*1] Uri Alon[1]
Pengfei Liu[1 2] Yiming Yang[1] Jamie Callan[1] Graham Neubig[1 2]
{luyug,amadaan,shuyanzh,ualon,pliu3,
yiming,callan,gneubig}@cs.cmu.edu



LLM

Code Interpreter

PYTHON INTERPRETER

Generate code as intermediate reasoning steps

Teach LLM how to decompose problems into coding steps

INPUT (with in-context example)

INPUT (with in-context example)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A:
#Roger started with 5 tennis balls.
tennis_balls = 5
# 2 cans of 3 tennis balls each is
bought_balls = 2 * 3 tennis balls.
# The answer is
answer = tennis_balls + bought_balls

Q: The Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

LLM OUTPUT

#The bakers started with 200 loaves
loaves_baked = 200
#They sold 93 in the morning and
# 39 in the afternoon
loaves_sold_morning = 93
loaves_sold_afternoon = 39
#The grocery store returned 6 loaves.
loaves_returned = 6
#The answer is
answer = loaves_baked - loaves_sold_morning
- loaves_sold_afternoon + loaves_returned

Code Interpreter

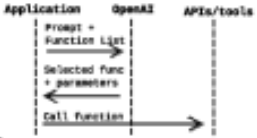
```
>>> print (answer)
74
```

ReAct in Practice

Structured Formats
JSON, XML, ...

```
```json
{
  "action": "Calculator",
  "action_input": "125 * 0.92"
}
```
```

Function Calling
Offload tool selection to API provider

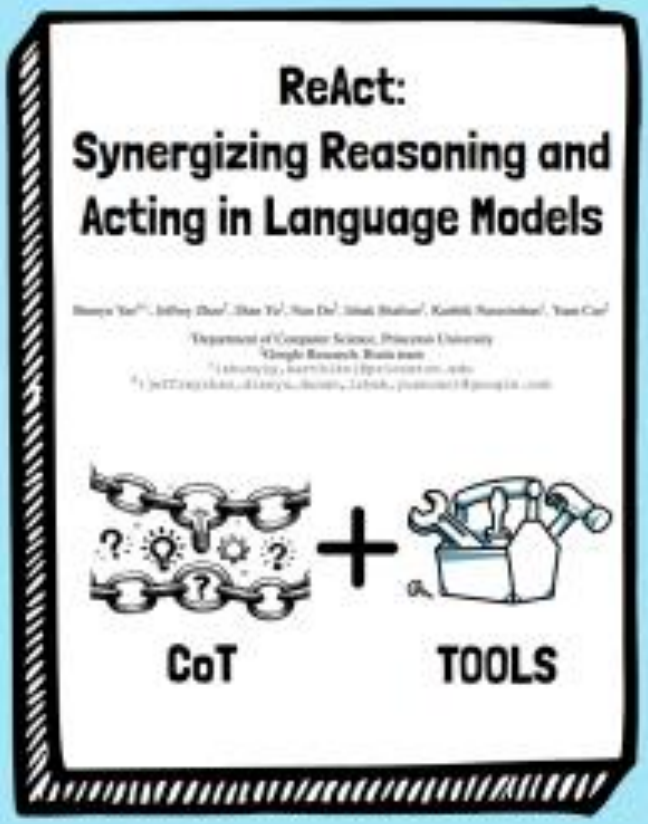


Fine-tuning
Improve tool usage and agentic workflows (e.g. InternLM)

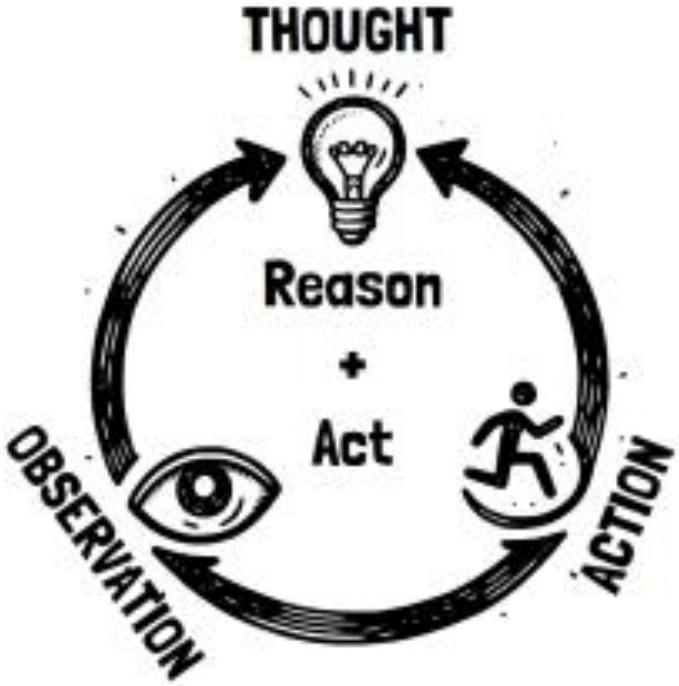
```
<|action_start|><|plugin|>
{
  "name": "calculator",
  "parameters":
    {"expr": "125*0.92"}
} <|action_end|>
```

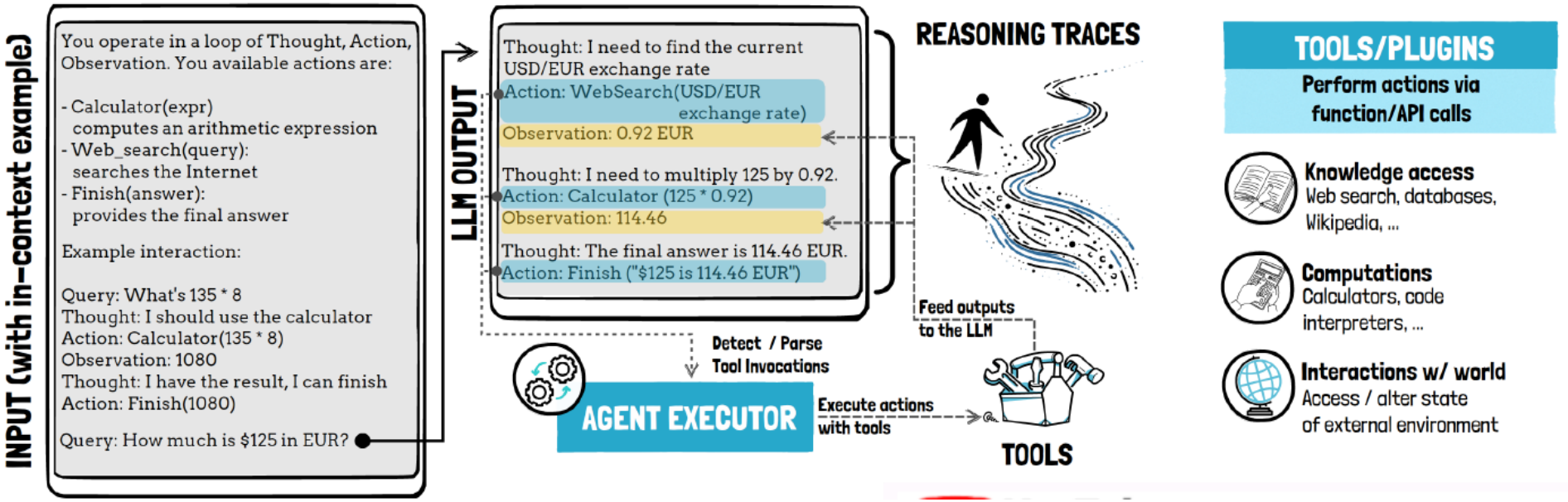
Modified ReAct
Adapted for specific workflows / use cases

GOOGLE, 2022



Use tools to solve problems




INPUT (with in-context example)
You operate in a loop of Thought, Action, Observation. You available actions are:
- Calculator(expr)
computes an arithmetic expression
- Web_search(query):
searches the Internet
- Finish(answer):
provides the final answer
Example interaction:
Query: What's 135 * 8
Thought: I should use the calculator
Action: Calculator(135 * 8)
Observation: 1080
Thought: I have the result, I can finish
Action: Finish(1080)
Query: How much is $125 in EUR?
LLM OUTPUT
Thought: I need to find the current USD/EUR exchange rate
Action: WebSearch(USD/EUR exchange rate)
Observation: 0.92 EUR
Thought: I need to multiply 125 by 0.92.
Action: Calculator (125 * 0.92)
Observation: 114.46
Thought: The final answer is 114.46 EUR.
Action: Finish ("$125 is 114.46 EUR")
REASONING TRACES
Feed outputs to the LLM
Detect / Parse Tool Invocations
AGENT EXECUTOR
Execute actions with tools
TOOLS
TOOLS/PLUGINS
Perform actions via function/API calls
Knowledge access
Web search, databases, Wikipedia, ...
Computations
Calculators, code interpreters, ...
Interactions w/ world
Access / alter state of external environment