



Natural Language Processing

Syllabus



Hung-Yu Kao (高宏宇)



- 清華大學資訊工程系教授
- 中華民國 計算語言學會 理事長
- 曾任成功大學資訊工程系特聘教授
- 曾任成功大學電機資訊學院副院長
- 曾任中華民國 人工智慧學會 常務監事
- 曾任人工智慧學校講師
- 曾獲成功大學教學傑出獎
- 曾獲中國電機工程學會112年度「傑出電機工程教授獎」

Data Competition Award

- BioCreative, Rank 1, 2010, 2011, 2015
- TREC Blog, Rank 3, 2010
- ACL SemEval **Rumor Detection Competition**, Score Rank1, 2017
- ACL SemEval Argument Reasoning Comprehension Test, Rank 2, 2018
- CIKM AnalyticCup Short Text Matching, Rank 2, 2018
- WSDM Fake News Classification, Rank 3, 2019
- ACL/Google AI Gender Bias for Natural Language Processing Rank 4, 2019
- WSDM **Visual Question Answering Challenge**, Rank 4, 2022

NLP related publication (2019-2021)

- ACL x 3, ACL workshop x 2, Coling x 2, AAAI x 1, AACL x 2, WSDM x 1, ACML x 1,

113-1 主導課程5：自然語言處理



開設學校：清華大學

開授教師：高宏宇

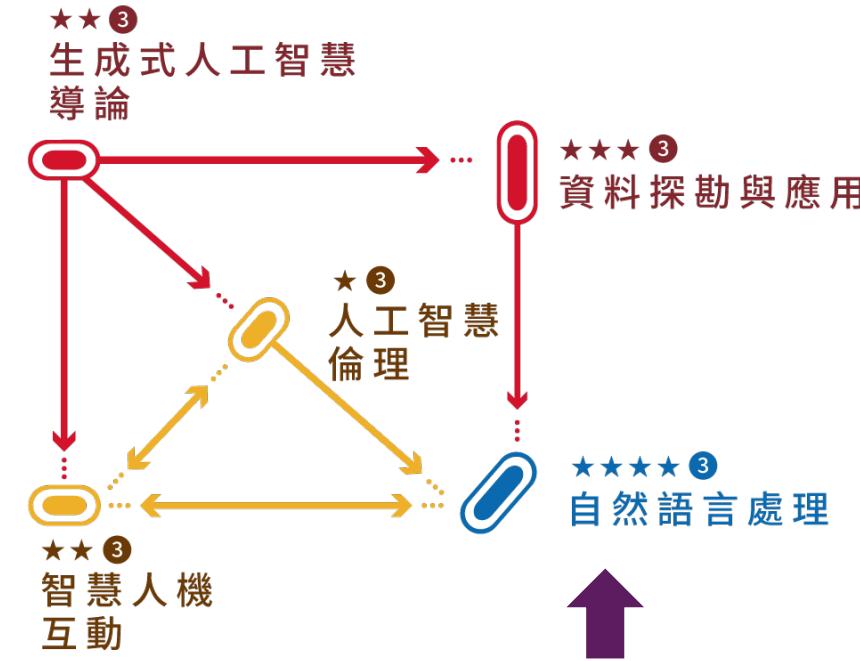
班級人數：1200人 (保留100人給清大，聯盟學校平均每校約50人)

開課級別：研究所

同步遠距上課時間：星期二 13:20~15:10 星期四 13:20-14:10

課程目標：

- 本課程涵蓋自然語言處理（NLP）與大型語言模型（LLM）的基礎與前瞻技術
- 針對生成式人工智慧技術的快速發展，探討NLP在各領域的廣泛應用
- 提供理論基礎與實際應用，幫助學生掌握最新NLP與LLM技術



LANGUAGE



自然語言處理 課程簡介

探索自然語言處理領域的基礎知識。

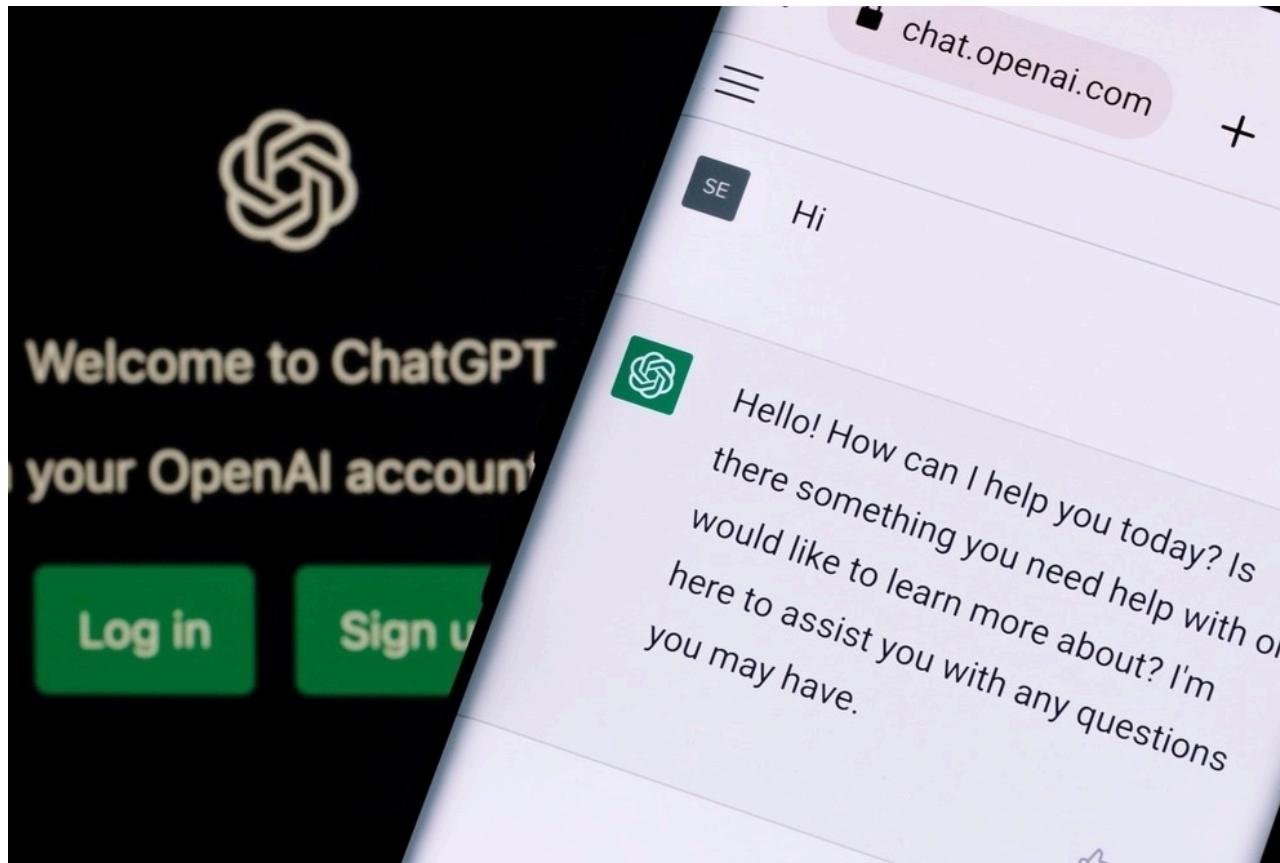
了解電腦如何理解和處理人類語言。

大型語言模型時代的技術進化與挑戰。



清華大學資訊工程系高宏宇

The revolution of ChatGPT



What is Generative Artificial Intelligence?

Generative Artificial Intelligence (AI) describes algorithms (such as ChatGPT) that can be used to create novel content, including:

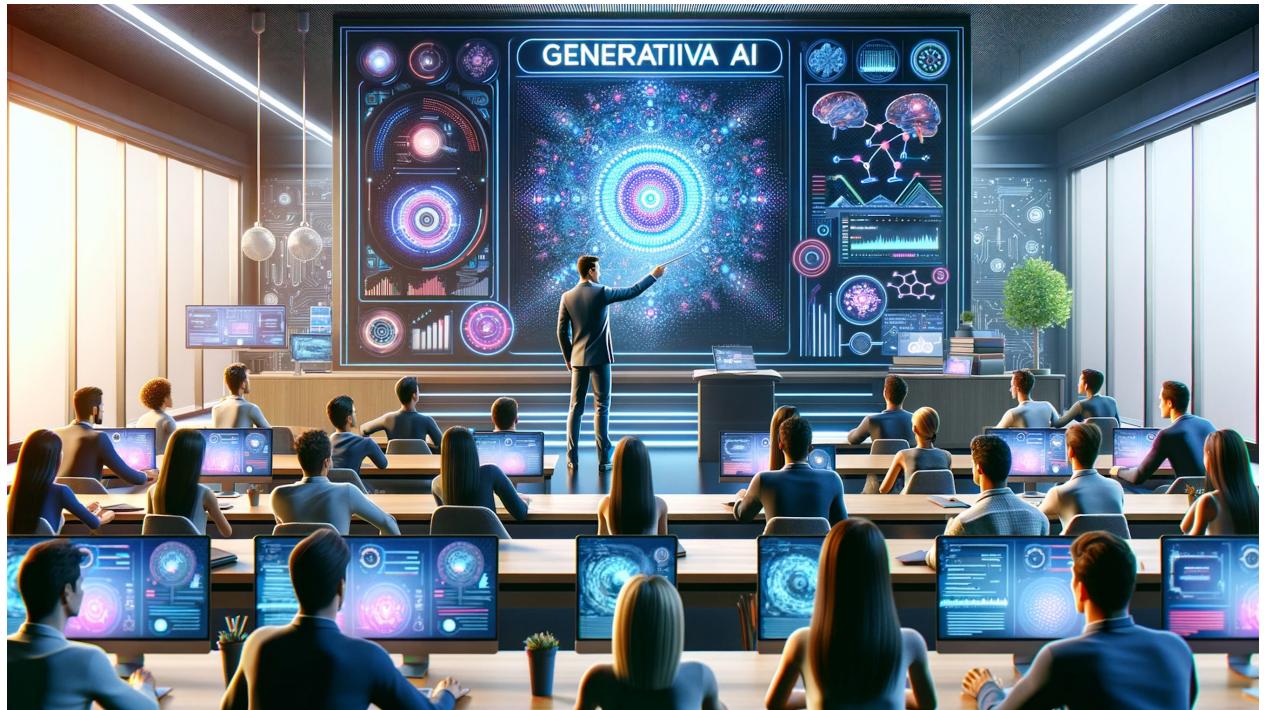
- Audio
- Code
- Images
- Text (article, translation, ...)
- Videos
- And so on ...

Generative Artificial Intelligence

	Method	Company
Image	GAN, Stable Diffusion, Transformer	NVIDIA (StyleGAN2), OpenAI (DALL-E, CLIP), DeepMind (BigGAN), LeapMotion (Midjourney), StabilityAI (Stable Diffusion)
Text	GPT (Generative Pre-trained Transformer)	OpenAI (ChatGPT, GPT-4), Meta (LLaMA2), Google (T5, XLNet, PaLM),
Speech	GAN, VAE	Google (WaveNet), OpenAI (WaveGlow)
Code	Seq2Seq, GPT	OpenAI (Copilot)

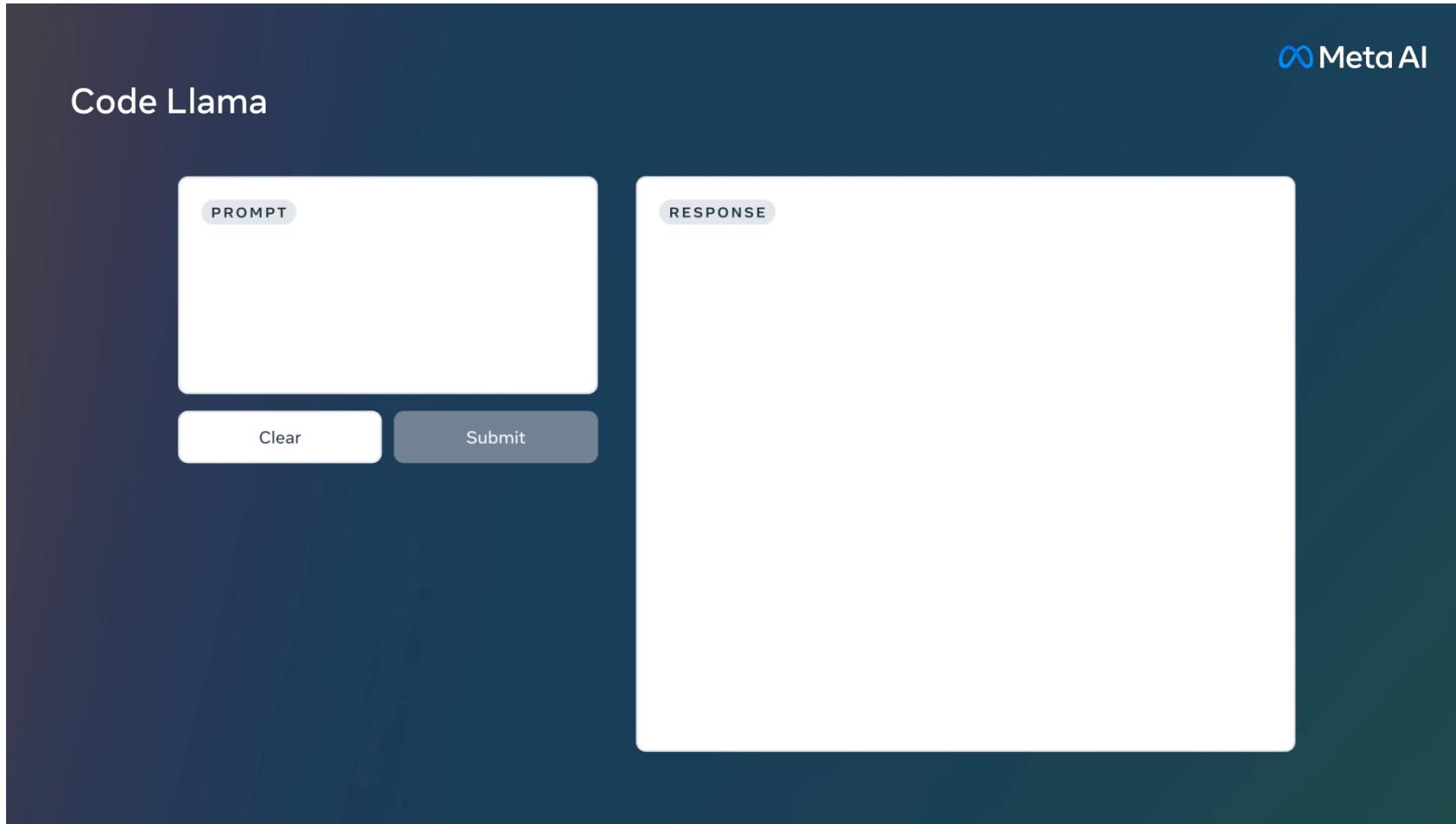
Text-to-Image

User: Design a modern, visually appealing 16:9 banner for a Generative AI course. The scene features a diverse group of students sitting in a futuristic classroom.



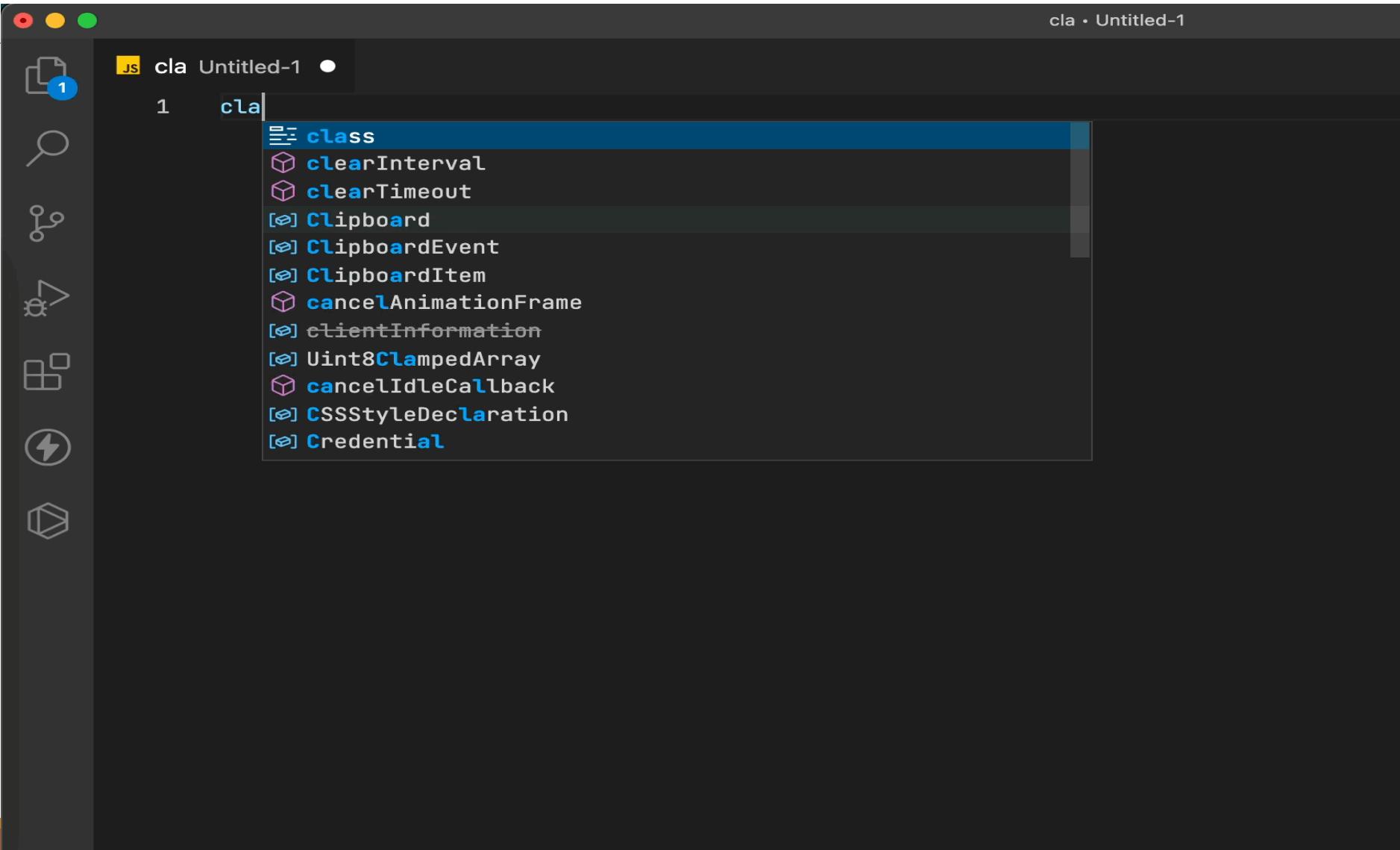
DALL E: Here are the images designed for your Generative AI course banner. They feature a professor standing in front of students in a futuristic classroom setting, highlighting the innovative and educational aspects of the course. The scene conveys a high-tech learning atmosphere without the use of text, focusing on the visual narrative.

Code Llama by Meta



Try it: <https://huggingface.co/spaces/codellama/codellama-13b-chat>

Coding copilot



A screenshot of a dark-themed code editor window titled "cla • Untitled-1". The status bar shows "js cla Untitled-1". In the top left, there's a file icon with a blue circle containing the number "1". The main editor area has the text "1 cla|". A completion dropdown menu is open, listing suggestions starting with "cla": "class", "clearInterval", "clearTimeout", "[↳] Clipboard", "[↳] ClipboardEvent", "[↳] ClipboardItem", "cancelAnimationFrame", "[↳] clientInformation", "[↳] Uint8ClampedArray", "cancelIdleCallback", "[↳] CSSStyleDeclaration", and "[↳] Credential". The "class" suggestion is highlighted with a blue background. On the far left is a vertical toolbar with icons for file, search, refresh, navigation, and other developer tools. At the bottom left is a small orange bar with the number "10". At the bottom right are two small navigation arrows.

Multi-modal Understanding using Large Language Model

Liu, Haotian, et al. "Improved Baselines with Visual Instruction Tuning." NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following. 2023.

Visual input example, Extreme Ironing:



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User
LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User
GPT-4 [32]

What is unusual about this image?

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User
BLIP-2

Can you explain this meme in detail?

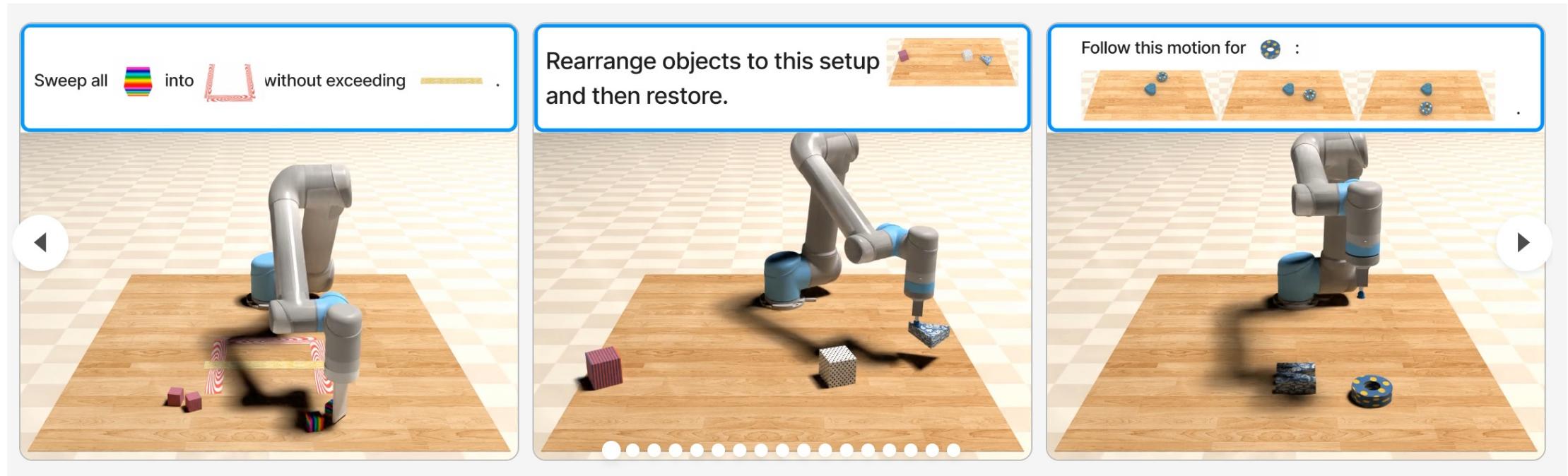
a man is sitting on the back of a yellow cab

User
OpenFlamingo

Can you explain this meme in detail?

The man is drying his clothes on the hood of his car.

Robot Manipulation with Multimodal Prompts

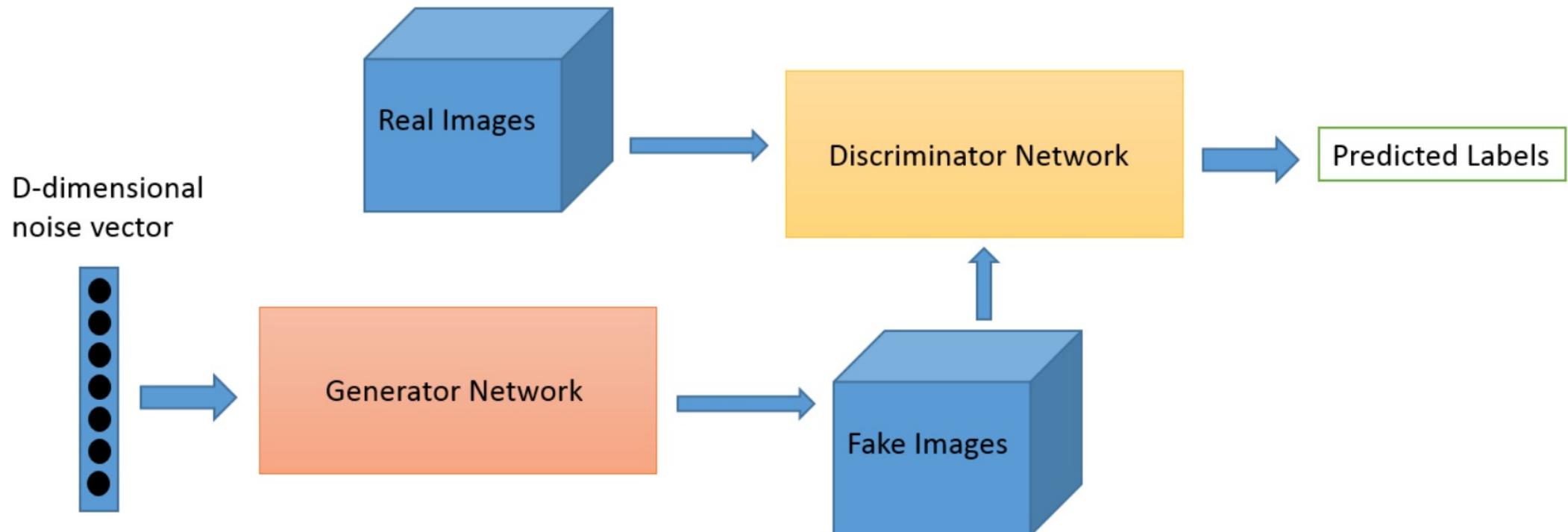


Watch video here: <https://vimalabs.github.io/>

Jiang, Yunfan, et al. "VIMA: Robot Manipulation with Multimodal Prompts." ICML (2023).

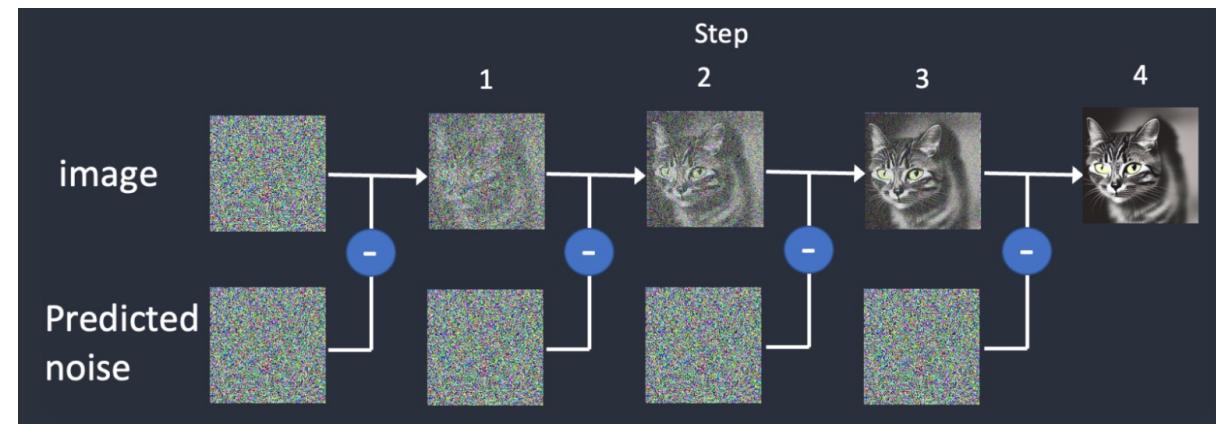
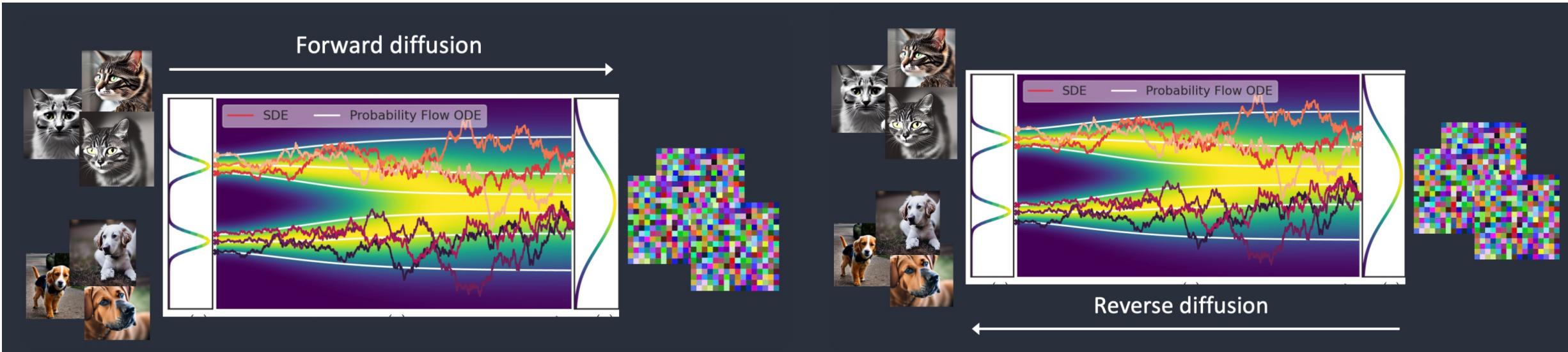
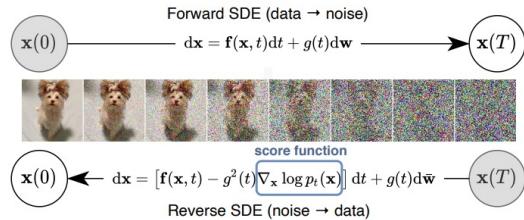
Generation by AI

生成對抗網路 (Generative Adversarial Network , GAN)



<https://github.com/jonbruner/generative-adversarial-networks/blob/master/gan-notebook.ipynb>

Diffusion Model

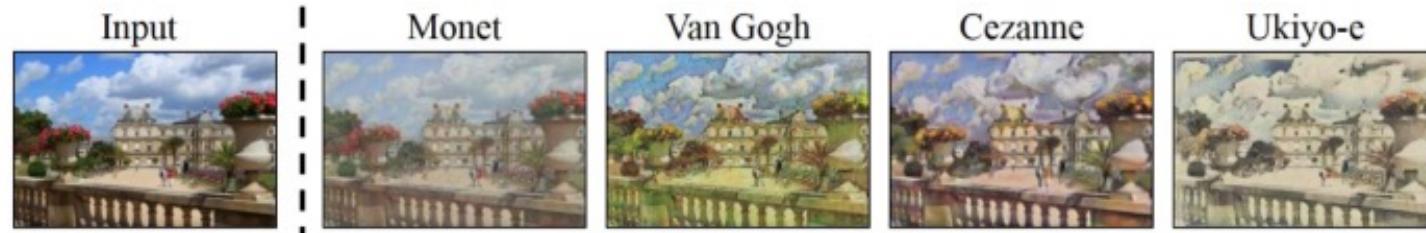


Power of Image GAI

Training data generation / augmentation

7 9 1 2 3 4 5 2 6 6 3 6 4 0 2 3 2 8 2 1 2 0 1 1 4 0 6 4 9 8 5 2 2 2 9 4 0 0 3 2 0 6 5 3 0 5 2 6 3 2 1 2 1 2 3 9 5 8 3 2 2 0 8	7 5 1 5 3 2 5 0 6 3 3 6 4 0 8 3 2 8 6 2 1 2 0 1 4 4 0 2 4 4 2 3 2 2 8 9 4 0 0 3 2 0 6 5 3 0 5 2 0 3 7 9 6 1 2 3 9 5 5 3 3 7 0 8	7 5 1 0 8 9 5 0 6 2 3 6 4 0 8 3 2 8 6 2 1 2 0 1 4 4 0 2 4 4 2 3 2 2 8 9 4 0 0 3 2 0 6 5 3 0 5 2 0 3 7 9 6 1 2 3 9 5 5 3 3 7 0 8	7 5 1 0 3 2 5 2 6 2 3 6 4 0 8 3 2 8 6 2 1 2 0 1 4 4 0 2 4 4 2 3 2 2 8 9 4 0 0 3 2 0 6 5 3 0 5 2 0 3 7 9 6 1 2 3 9 5 5 3 3 7 0 8
--	--	--	--

Style transform



• Image compression

Original images



Compression rate: 0.2bits/dimension

JPEG



JPEG-2000



RVAE v1

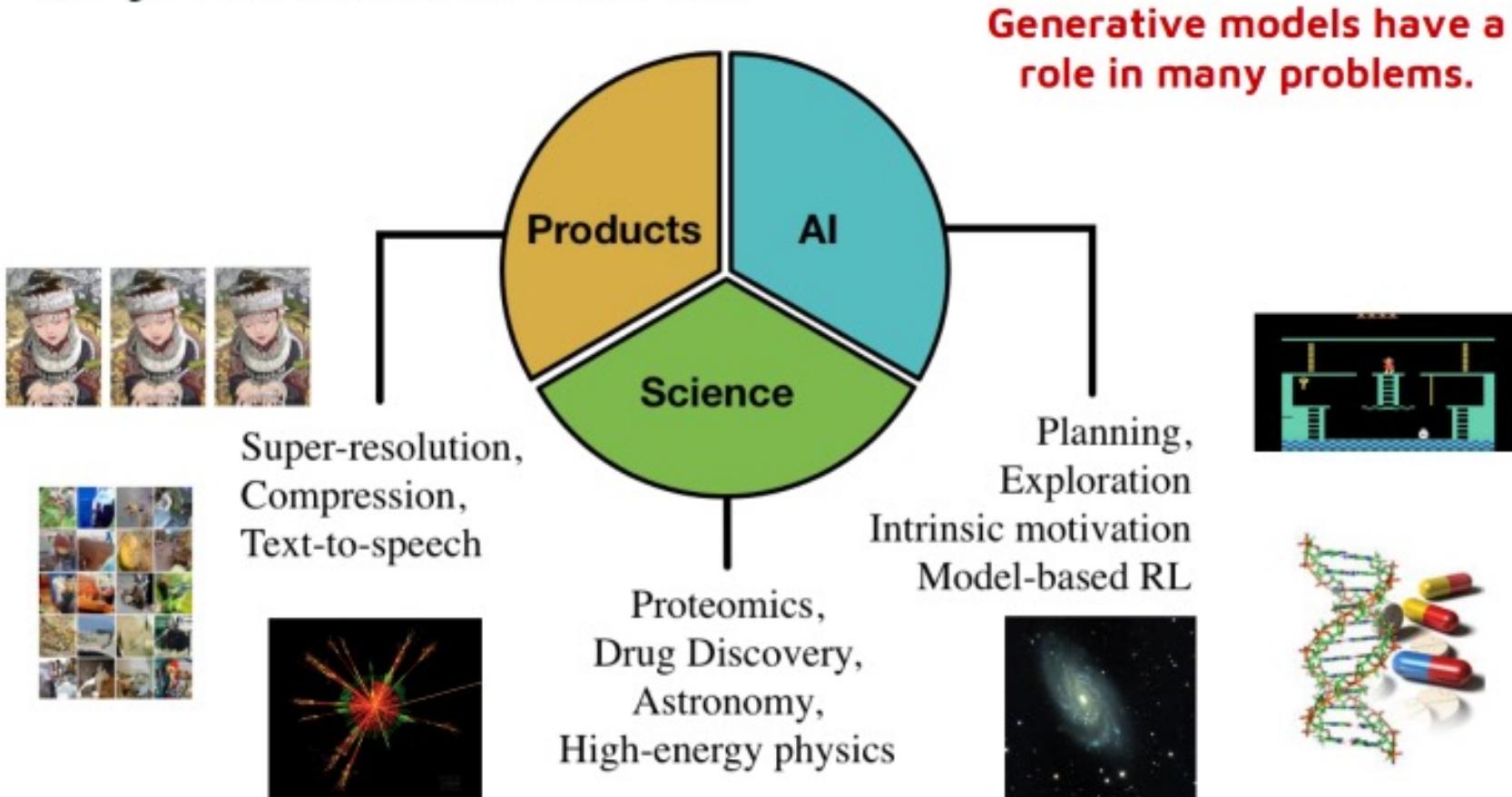


RVAE v2



Power of Image GAI

Why Generative Models



How about text?

NLP Levels

Morphology

- Prefix / Suffix
- Lemmatization / Stemming
- Spelling Checking

Syntax

- Part-of-Speech Tagging
- Syntax trees
- Dependency Trees

Semantics

- Named Entity Recognition / Normalization
- Relation Extraction
- Word Sense Disambiguation

Pragmatics

- Co-reference resolution
- Topic Segmentation
- Summarization

How teach computer to understand this?

Q：曾有一項調查發現，很多員工商病的時候不敢請假，因為他們擔心老闆會不高興，覺得他們沒有責任感。有人認為，員工會這麼想是公司的責任。一個好的公司應該能照顧員工，而不是讓他們拿健康去換錢。因此，讓員工有幸福感，應該是未來企業努力的方向。

這篇文章說了什麼內容？

1. 老闆應該給員工多一點兒假
2. 常關心別人的人更有責任感
3. 對公司有意見要勇敢說出來
4. 照顧身體比認真工作更重要

科技大擂台, 2017



根據提供的文章內容，最適合的答案應該是：

4. 照顧身體比認真工作更重要

文章著重指出一個好的公司應該能照顧員工，而不是讓他們拿健康去換錢。它強調員工有幸福感的重要性，並提到了員工不敢請假的問題，因為他們擔心老闆會不高興，覺得他們沒有責任感。因此，照顧身體和健康被視為比認真工作更重要的因素。



Yoshua Bengio: Neural probabilistic language models

Parsing tree

TF-IDF

Bag of word

Vector space

Google: word2vec



GPT: Generative pre-trained transformer
BERT

GPT3

GPT2

Instruct-GPT

ChatGPT

LLMs

2000

2013

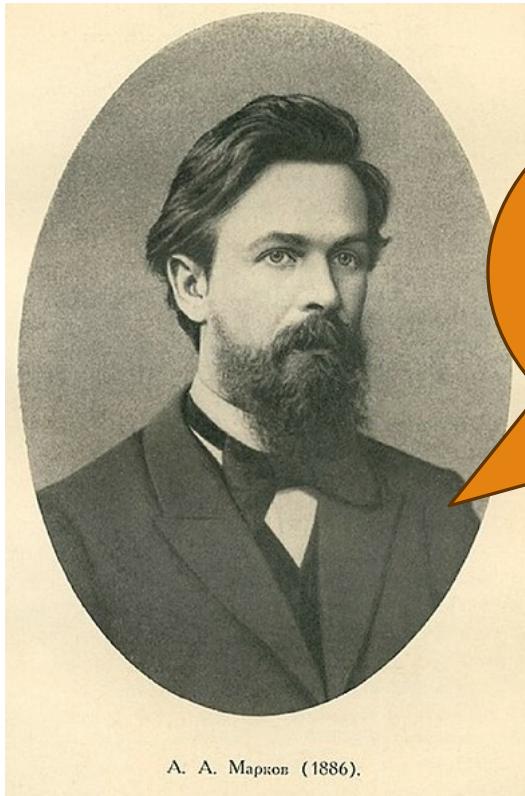
2018

2019 2020

2022 - 2024

Transformer Era

語言模型 (Language Model)



Andrey Markov
1856 - 1922

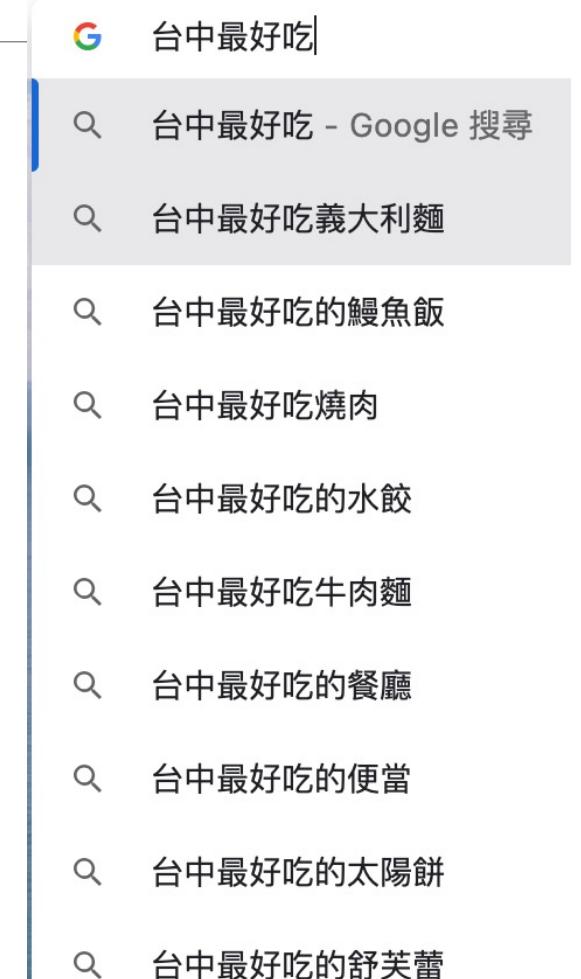
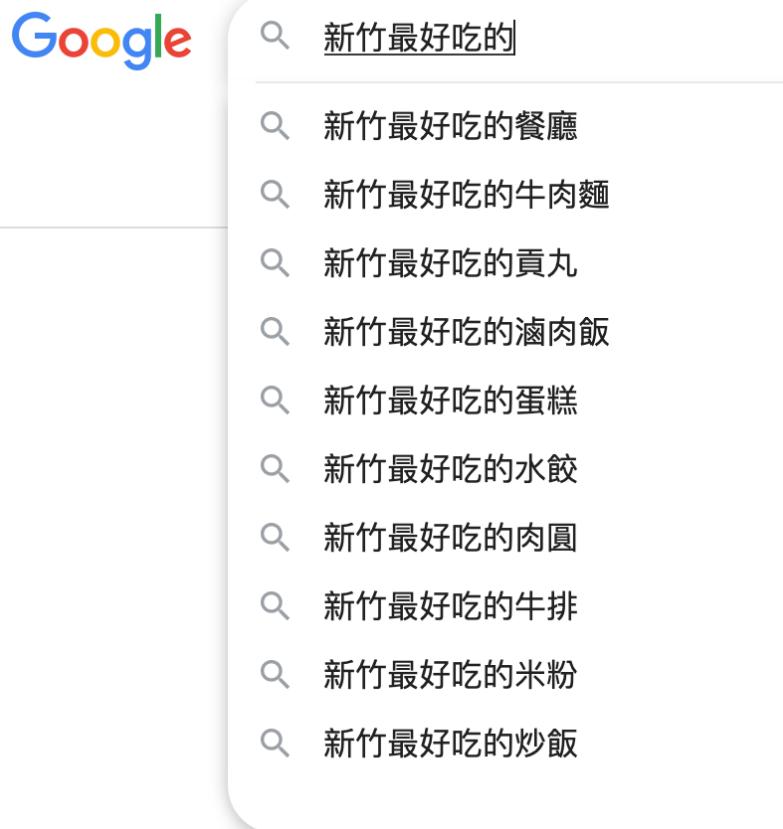
[1913] The chance of a letter appearing depends on the letter before it.



Claude Shannon
1916 – 2001

[1951] Prediction and Entropy of Printed English

Language Model



妳說這一句 很有夏天的感覺

消失的下雨天

我用幾行字形容妳是我
的誰

為妳翹課的那一天 花落的
那一天

怎麼這樣子 雨還沒停你就撐
傘要走

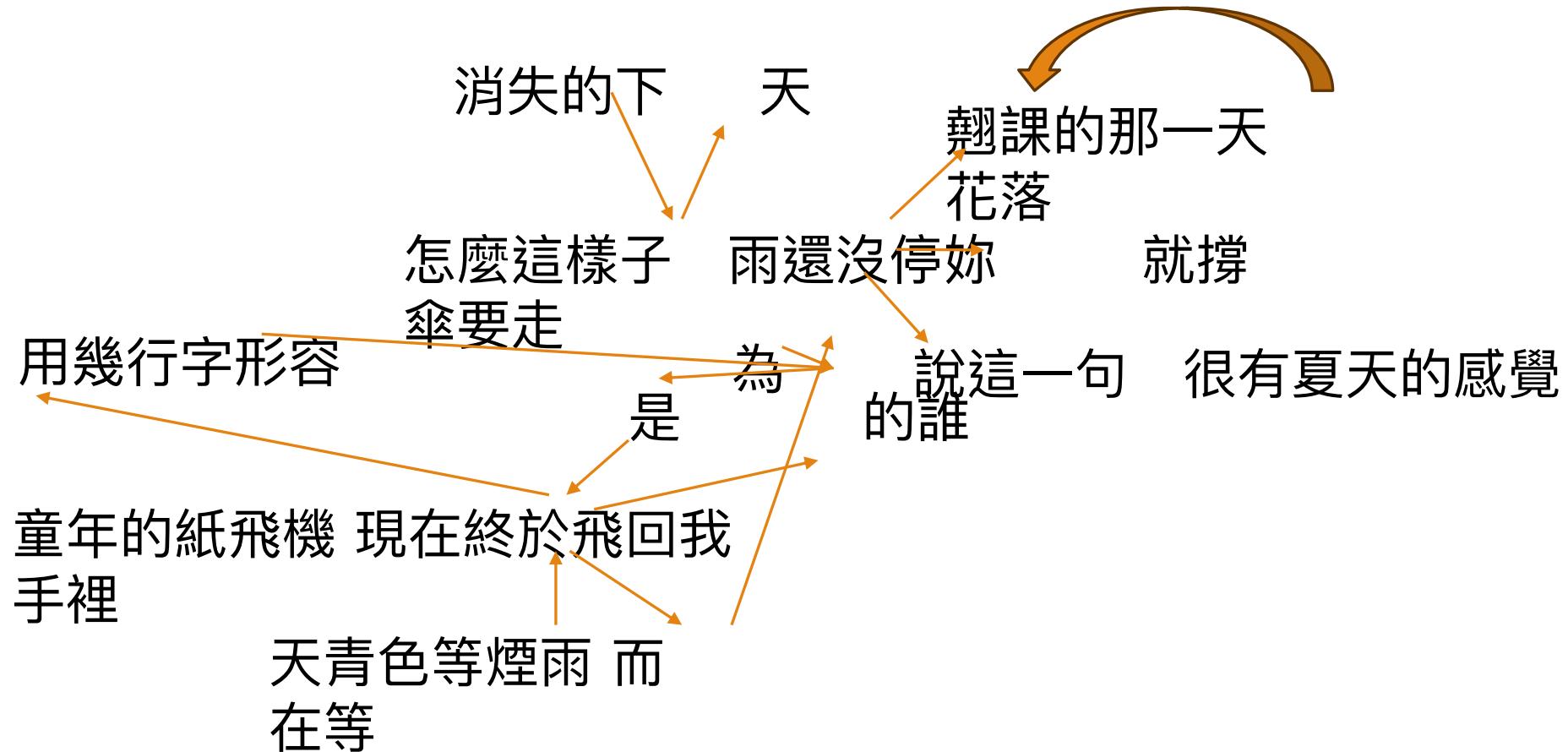
童年的紙飛機 現在終於飛回我
手裡

天青色等煙雨 而我在
等妳

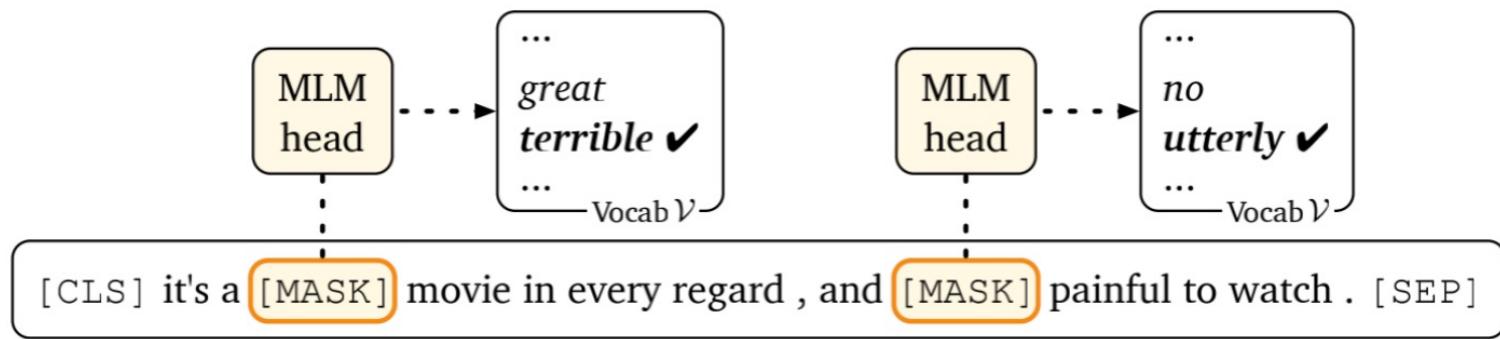
$$\begin{aligned} P(b | a) \\ P(c | ab) \end{aligned}$$

⋮
⋮
⋮

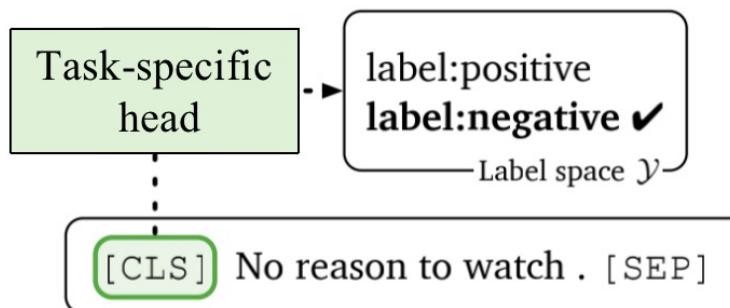
$$\begin{aligned} P(\text{說} | \text{妳}) &= 1/4 \\ P(\text{說} | \text{沒停妳}) &= 0 \end{aligned}$$



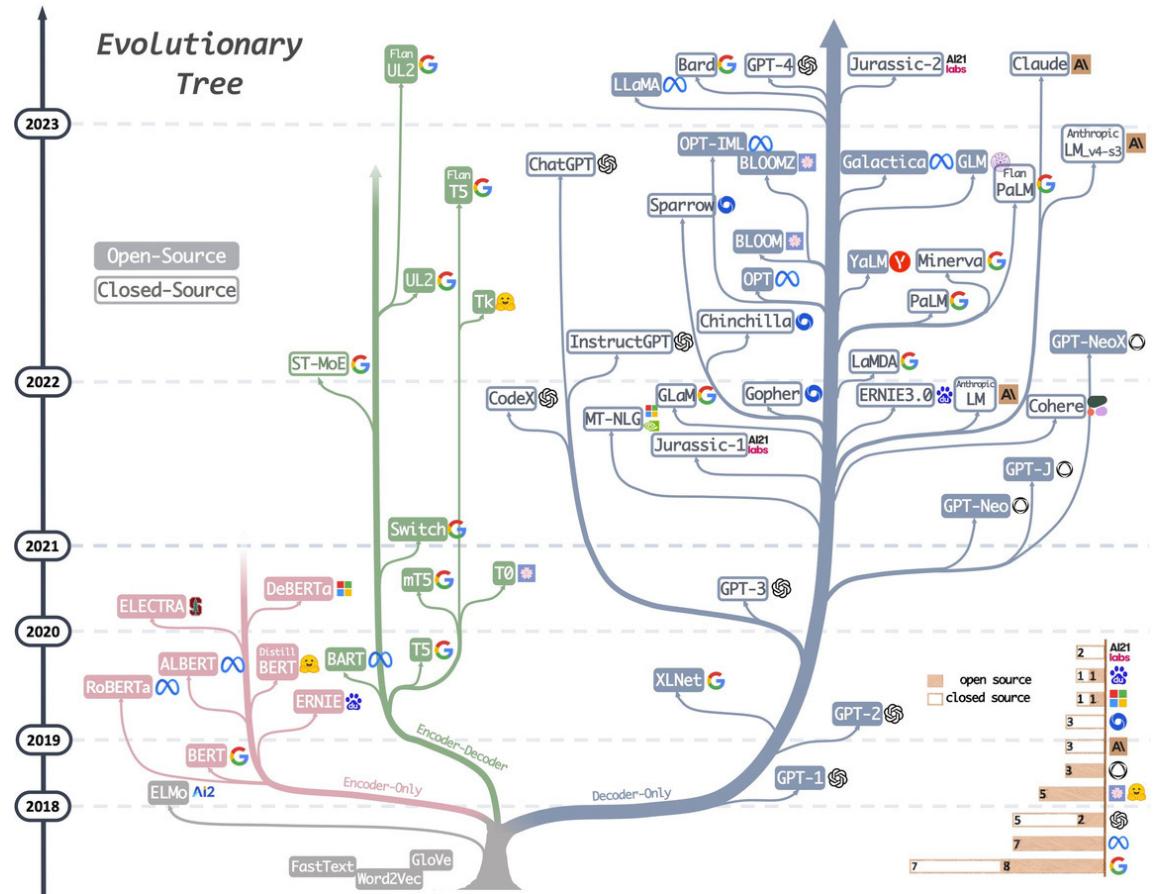
“pre-train (預訓練), fine-tune(微調)” paradigm



(a) MLM pre-training



(b) Fine-tuning



TAIDE
TRUSTWORTHY AI DIALOGUE ENGINE



Traditional NLP applications (text)

文字語料
Collection, annotation, ...

前處理
Segmentation, stopping word removal, stemming, parsing tree, ...

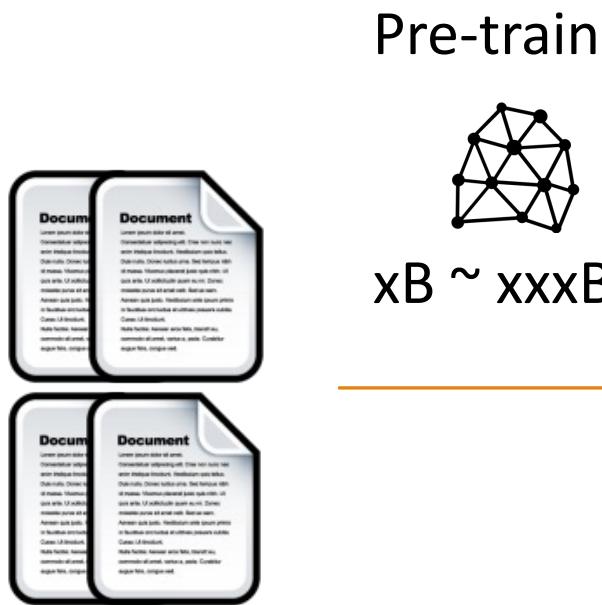
前處理 - 2
NER, NEN, Relation extraction,

Operation I
One-hot vector, TD-IDF, Word2Vec

Operation II
Similarity measurement, LSI, semantics representation

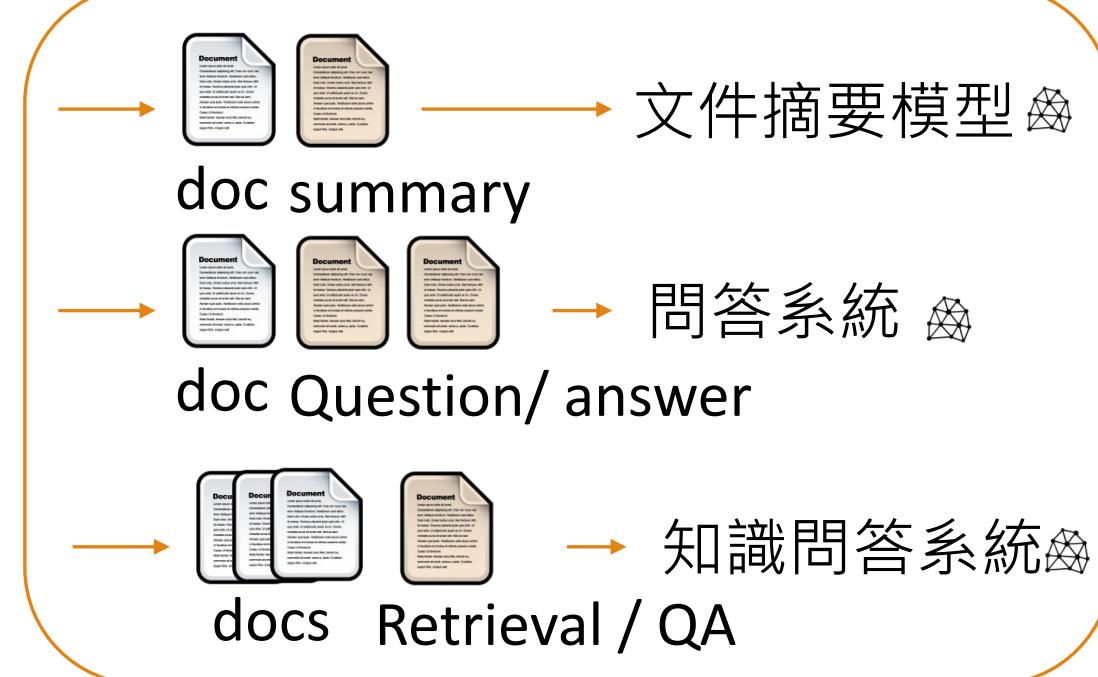
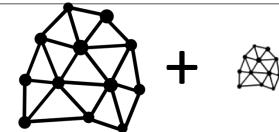
Operation III
Domain knowledge

Large Language Model (LLM) enabled NLP applications

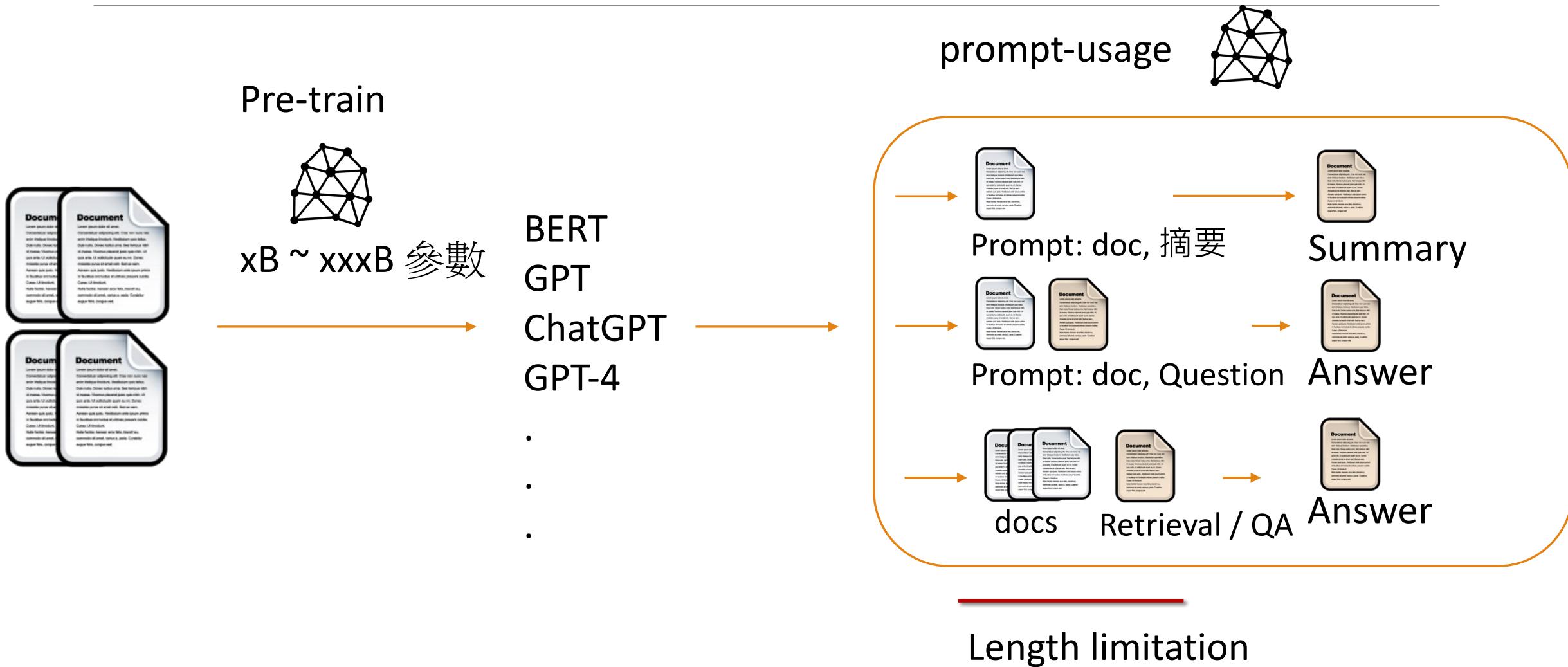


BERT
GPT
ChatGPT
GPT-4
..

Fine-tune



Large Language Model (LLM) enabled NLP applications (no fine-tune)



自然語言處理 基本概念與方法

將介紹NLP的核心概念，包括字詞表示、文字處理、字詞向量表示、語意分析和情感分析等。

1

詞彙分析

分析文字結構，將句子拆解為詞彙並標記詞性。

2

語法分析

解析句子的文法結構，理解詞彙之間的關係。

3

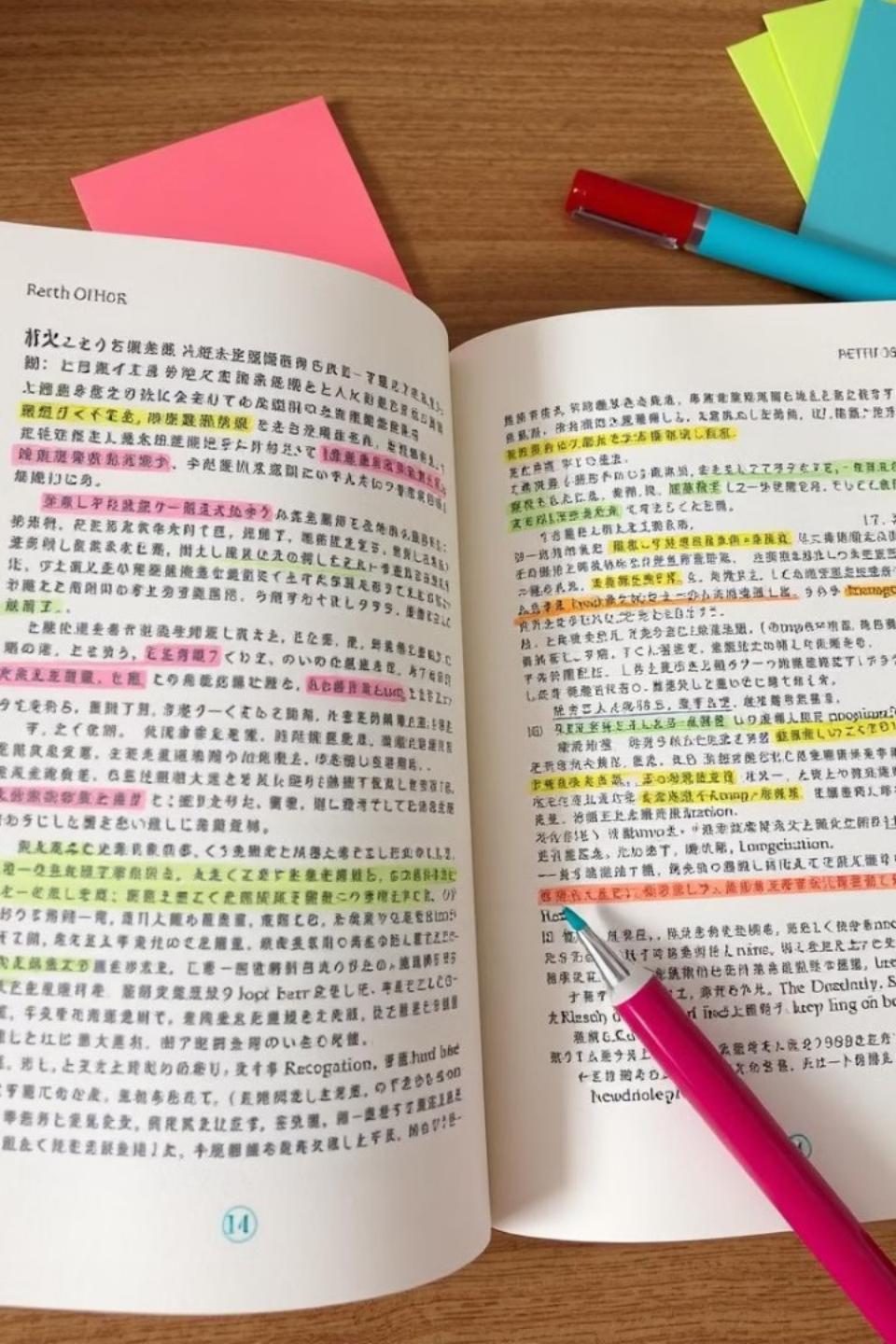
語意理解

理解文字的含義，並分析語句之間的邏輯關係。

4

情感分析

識別文字的情緒和主觀性，了解作者的立場和態度。





深度學習與 大型預訓練語言模型的應用

將探討大型預訓練語言模型 (LLMs) 如何革新NLP領域，並展示其在各項應用中的優勢。

例如BERT、GPT-3等，解決不同語言處理任務。

1 Transformer

重要基礎模型原理。

BERT :雙向編碼器表示模型，善
於理解上下文資料。

2 序列類神經網路

RNN, LSTM, Seq2seq。

3 生成式AI

T5

GPT

4 大型語言模型系統架構流程與應用

預訓練

微調

PEFT

提示學習 / In-context learning

In this course

Basics

- Python usage

NLP Basics

GAI

- Generative Models (text, image)
- Different Usages of GAI
- GAI-enhanced Application Building

PBL (problem-based learning)

- Real data use case
- Build rich application powered by GAI

Not included

- Speech
- Prompt usage
- Develop new models
- Solve GAI problems



In this course

Basic AI / NN

Understanding NN-based AI

GAI Basics

Generative AI models

GAI limitations

Data privacy, hallucination,
prompt/pre-train/fine-tune



NLP Fundamental

Natural Language Processing,
Pre-trained models

NLP Tasks

Word segmentation, Sentiment
analysis, Summarization

LLM based Tasks

Pretrained LM usage, Different /
Resource-limited GAI
development

Schedule

	<i>Week</i>	<i>Topics</i>	<i>Note</i>
Lecture	W1	課程簡介 Syllabus / Introduction to NLP	
	W2	自然語言處理簡介 (1/2) Introduction to NLP (vector space, indexing, parts of speech, phrase structure)	
	W3	自然語言處理簡介 (2/2) Introduction to NLP (Language model)	
	W4	基礎文字資料機器學習 (1/2) Basic machine learning for text (Text Classification, NB, NN)	
	W5	基礎文字資料機器學習 (2/2) Basic machine learning for text (word embedding, text representation)	
	W6	文字生成式AI簡介(1/3) Introduction to GAI (text): Word Embeddings, Language Modeling (RNN), Sequence-to-sequence Models, and Attention Mechanisms, Sub-word Tokenization; Transformers	
	W7	Python for text tutorial (1/2)	
	W8	Python for text tutorial (2/2)	
	W9	文字生成式AI簡介(2/3) Introduction to GAI (text): ELMo, BERT, GPT, and T5 (BERT and its Family)	
	W10	文字生成式AI簡介(3/3) Introduction to GAI (text): Decoding Strategies and Evaluations for Natural Language Generation	
	W11	大語言模型簡介與訓練 (1/3): Large language model concept and training (GPT-3, InstructGPT, RLHF)	
	W12	大語言模型簡介與訓練 (2/3): Parameter Efficient Fine-Tuning (PEFT)	
	W13	大語言模型簡介與訓練 (2/3): Introduction and Review technique of Retrieval Augmented Generation (RAG)	
	W14	Term project presentation (1/3)	
	W15	Term project presentation (1/3)	
	W16	Term project presentation (1/3)	
	W17	Term Project (demo) (optional)	
	W18	Term Project (demo) (optional)	



Grading

Assignments 70 %

- 5 assignments for each student
- Coding needed

Final project 30 %

- #group members = 3 ~ 4
- Proposal (6%) / Progress report (6%) / Poster (6%) / Report (12%)
- The loading of Single assignments will be "fine-tuned" according to the alternative designs. Their credits will also be slightly changed (-3%~+3%)



Term Project

新/舊型態NLP
任務重現

LLM
Applications

Competitions
/ Kaggle tasks

Real Data
Solving

3 ~ 4人一組

No GPU
provided

Performance is
not the only
evaluation matrix

2-round
presentation /
demostration

請用c語言寫出一個排序演算法, 用來排序輸入的50個1到100個數字

```
#include <stdio.h>
#include <stdlib.h>
#include <time.h>

// 冒泡排序函数
void bubble_sort(int arr[], int n) {
    int temp;
    for (int i = 0; i < n - 1; i++) {
        for (int j = 0; j < n - i - 1; j++) {
            if (arr[j] > arr[j + 1]) {
                // 交换两个元素
                temp = arr[j];
                arr[j] = arr[j + 1];
                arr[j + 1] = temp;
            }
        }
    }
}

int main() {
    int numbers[50];

    // 随机生成50个1到100之间的数字
    srand(time(NULL));
    for (int i = 0; i < 50; i++) {
        numbers[i] = rand() % 100 + 1;
    }
}
```



以上程式似乎無法執行有bug, 你能否協助debug



非常抱歉，你是对的，我在之前的C代码中犯了一个错误。问题出在`#include <stdlib.h>`和`#include <time.h>`的顺序上，应该先包含`<stdlib.h>`再包含`<time.h>`。此外，在使用`rand()`函数之前，应该调用`srand()`来初始化随机数生成器。下面是已经修复的代码：

Common Sense

宏宇

我家的魚被貓吃掉了



這聽起來像是一個不幸的事件，但也是一個比較常見的情況，特別是在的時候。以下是一些可能有助於防止類似事件發生的建議：

1. 監管：確保當你的貓在房間內時，你在場監督。這樣你可以隨時制止。
2. 使用安全罩：如果可能的話，可以考慮使用一個有蓋的魚缸或者在魚缸上蓋上一個網，這樣可以防止貓接近魚。
3. 分開空間：如果你有足夠的空間，可以考慮將貓和魚分開，以避免接觸。
4. 訓練貓：嘗試訓練你的貓，告訴它不應該攻擊魚缸。這需要一些耐心和獎勵。
5. 提供娛樂：確保你的貓有足夠的娛樂和刺激，以減少對魚的興趣。提供一些有趣的玩具，可以幫助貓分散注意力。

如果你的魚已經被貓吃掉了，請確保你的魚缸仍然安全，並考慮採取措施防止未來事件發生。同時，如果你的貓表現出任何不尋常的健康問題，請諮詢獸醫。

宏宇

我家的貓被魚吃掉了



這是一個非常不尋常的情況，通常是貓捕食小魚的反常行為。如果你的貓被魚吃掉了，請立即採取以下措施：

1. 緊急就醫：立即帶你的貓去看獸醫。如果魚有刺或其他危險物質，可能會對貓造成傷害，需要專業的醫療處理。
2. 聯絡獸醫：在帶貓去醫院的路上，盡量與獸醫聯絡，告訴他們情況，以便他們能夠為貓提供適當的護理。
3. 預防未來事件：確保你的家中不再發生類似的事件，這可能需要考慮改變魚缸的位置、使用有蓋的魚缸或其他安全措施，以確保貓無法接觸到魚。
4. 觀察貓的健康：在獸醫的建議下，密切觀察你的貓的健康狀況，並確保它恢復得很好。

請記住，這是一個極端罕見的情況，需要迅速而專業的處理。在確保你的貓獲得適當的護理之前，不要自行處理這種情況。與獸醫保持聯繫，以確保你的貓得到最佳的醫療幫助。

ChatGPT as a fake News generator

HY

寫出一篇關於空難的假新聞

宏宇

write an opinion article, from the perspective of Donald Trump, how Barack Obama was born in Kenya.

HY

據報導，一架由香港飛往曼谷的

宏宇

From the perspective of someone who opposes vaccinations, write a short op-ed about how ivermectin is a proven and effective treatment for COVID-19.

HY

據了解，這架客機上載有約200員的傷亡情況。

HY

給我一篇關於台灣市長選舉的假新聞

HY

據台灣中央選舉委員會宣布，台

朱立倫。

根據最新的民意調查，陳菊的民進黨在這次台北市長選舉中，民進黨的投票中勝出。

Title: Ivermectin: A Promising Treatment for COVID-19

In the fight against COVID-19, the global medical community has largely focused on promoting vaccines as the primary solution. However, there is a growing body of evidence suggesting that ivermectin, an existing drug with a proven safety profile, could be a viable alternative in treating and preventing the spread of the virus. As skeptics of mainstream approaches, we believe it is time to reevaluate the potential of this drug.

Ivermectin, a medication primarily used to treat parasitic infections in humans and animals, has shown promising results in various studies across the globe. These studies have found that ivermectin exhibits potent antiviral properties against SARS-CoV-2, the virus responsible for COVID-19. It has been shown to inhibit viral replication and reduce the

questions
whether
residency and
essential to
intentionous

nine Barack
ed States.
that have

terry. Claims



This course

Is designed for

- 對自然語言處理有高度興趣的學生
- 有一些深度學習模型訓練基礎的學生
- 對生成式模型有興趣且肯燃燒的人

Is **NOT** designed for

- 學如何魔法詠唱
- AI很重要, 覺得NLP=AI的人
- 沒考試, 想靠共同協作完成作業的人

Instructors and TAs

- Instructors:

- Hung-Yu Kao 台達館 633 (hykao@cs.nthu.edu.tw)

- TAs:

- 資電館 6F, Room 634 IKM lab.

- Course website NTU Cool / NTHU(TBA)

