

GPT-4 來了!  
有甚麼特別的地方嗎？

---

# GPT-4 Technical Report

---

OpenAI\*

## Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

<https://openai.com/research/gpt-4>



# What do we know?

## **2 Scope and Limitations of this Technical Report**

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [33] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [34]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

# GPT 看得見了!

User Can you explain why this is funny. Think about it step-by-step.



該功能目前  
無法測試

GPT-4 The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such

# 有人說 ChatGPT 其實會讀圖片連結？真的嗎？



<https://speech.ee.ntu.edu.tw/~hylee/ml/man.png> 你看到甚麼

這個圖片連結並不存在!



根據圖片，我看到的是一個人的手拿著一個物體，該物體似乎是一個手槍。這個手槍看起來比較老舊，可能是一把經典的手槍。



<https://speech.ee.ntu.edu.tw/~hylee/ml/girl.png> 你看到甚麼



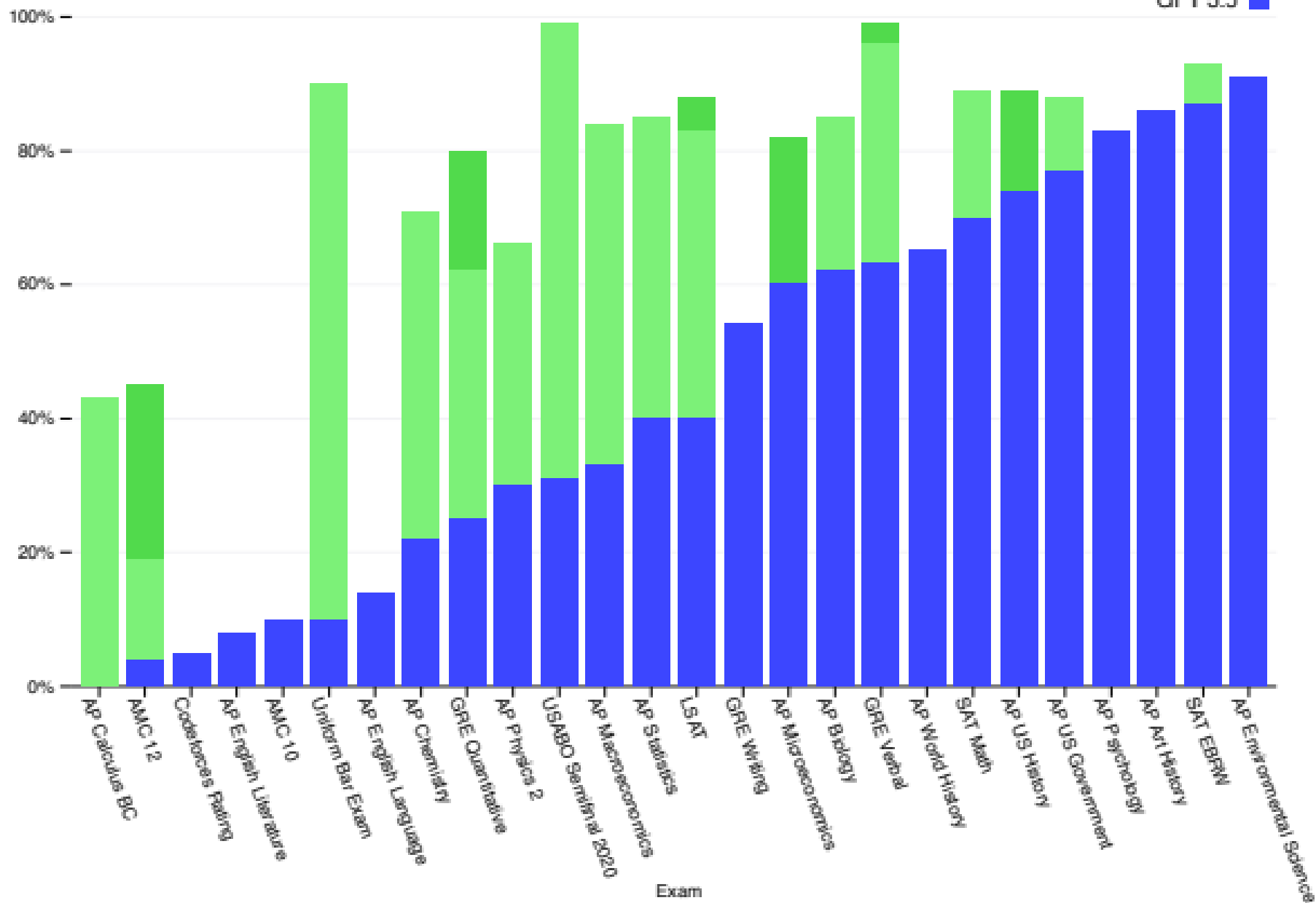
作为一个AI模型，我无法真正“看到”像人类一样看到的图像，但是我可以对这张图像进行分析和描述。

根据我所接受的信息，这张图像似乎是一张漫画或插画，它描绘了一个穿着校服的女孩子。女孩子的头发很长，披在肩上，她的脸朝向观察者，眼睛很大，嘴巴微微张开，似乎正在微笑或说话。她的手臂略微弯曲，手里拿着一本书。她的身体姿势向左倾斜，似乎在行走中。背景是一些简单的图案，可能是墙壁或纹理。

## Exam results (ordered by GPT 3.5 performance)

Estimated percentile lower bound (among test takers)

GPT 4 ■  
 GPT 4 (no vision) ■  
 GPT 3.5 ■





雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？



假设雞有 $x$ 隻，鴨有 $y$ 隻，兔有 $z$ 隻。

1.  $x + y + z = 30$  (雞、鴨、兔共30隻)
2.  $2x = y$  (雞的數量是鴨的2倍)
3.  $2x + 2y + 4z = 72$  (雞和鴨有2條腿，兔有4條腿)

我們可以用第二個方程解出

$$y = 2x$$

將 $y = 2x$ 代入第一個方程，

$$x + 2x + z = 30$$



雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？



根据题目，我们可以

1. 鸡 + 鸭 + 兔 = 30
2. 2 \* 鸡的数量 = 鸭的数量
3. 2 \* 鸡 + 2 \* 鸭 + 4 \* 兔 = 72

首先将方程1和方程2

然后，我们将方程3

化简方程3：2 \* 鸡 +

接着，我们将两个方



雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？



根据题目描述，我们可以用以下方程组表示这个问题：

设鸡的数量为 $x$ ，鸭的数量为 $y$ ，兔的数量为 $z$ ，那么：

$$x + y + z = 30$$

$$2x = y$$

$$4x + 2y + 4z = 72$$

我们可以先通过第二个方程解出 $y$ ：

$$y = 2x$$

然后将 $y$ 代入第一个方程：

$$x + 2x + z = 30$$

$$3x + z = 30$$

最後還是算錯了...



雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？

假設雞的數量為 $x$ ，鴨的數量為 $y$ ，兔的數量為 $z$ 。

1.  $x + y + z = 30$  (總數量)
2.  $2x = y$  (雞的數量是鴨的2倍)
3.  $2x + 2y + 4z = 72$  (總腿數)

根據第二個條件，我

將這個方程代入第一

$$x + 2x + z = 30$$
$$3x + z = 30$$

雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？

假設雞的數量為 $x$ ，鴨的數量為 $y$ ，兔的數量為 $z$ 。

因為題目告訴我們「雞的數量是鴨的2倍」，所以可以列出以下等式：

$$x = 2y$$

同時，三種動物的總數為30，所以可以列出以下等式：

$$x + y + z = 30$$

另外，三種動物的腳的總數為72，因為每隻雞有2隻腳、每隻鴨有2隻腳、每隻兔有4隻腳，所以可以列出以下等式：

$$2x + 2y + 4z = 72$$

雞、鴨、兔共30隻，72條腿。其中雞的數量是鴨的2倍，那麼雞有幾隻？

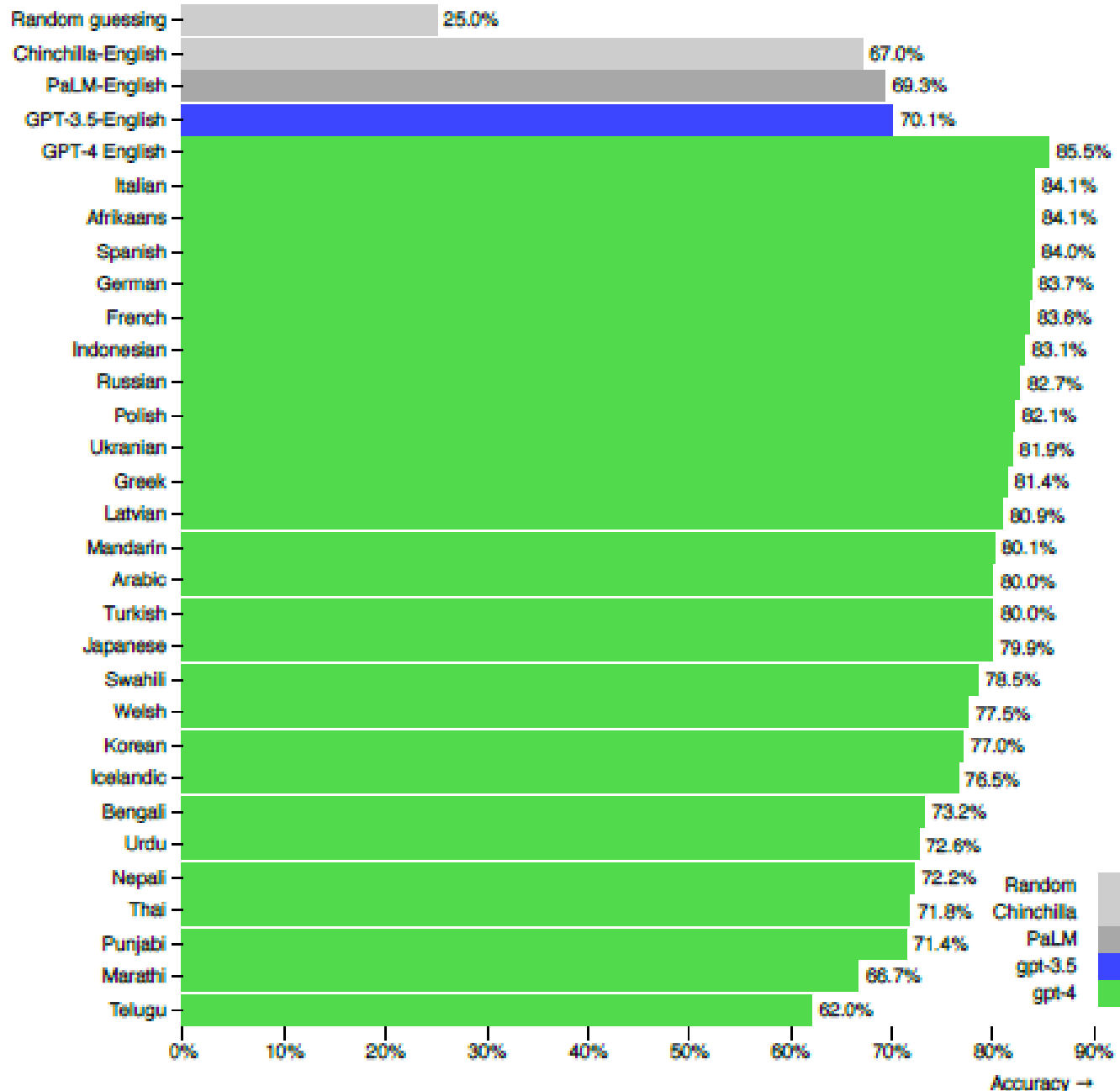
設鴨的數量為 $x$ ，則雞的數量為 $2x$ ，兔的數量為 $30 - x - 2x = 30 - 3x$ 。

由於每隻雞有2隻腿、每隻鴨有2隻腿、每隻兔有4隻腿，因此總共有的腿數為：

$$2(2x) + 2x + 4(30 - 3x) = 72$$

化簡後得到  $x = 6$ ，即鴨的數量為 6 隻，雞的數量為  $2x = 12$  隻。因此，雞有 12 隻。

## GPT-4 3-shot accuracy on MMLU across languages





Tsò-lâng nā pîn-tuānn, tsit-sì-lâng bē khuìnn-uáh. 做人若貧憐，一世人袂快活



Tsò-lâng nā pîn-tuānn, tsit-sì-lâng bē khuìnn-uáh.



I apologize, but I am not able to understand the language of the text you have provided. It appears to be written in a non-Latin script which I am not programmed to read. If you can provide a translation or context, I will be happy to assist you.



Tsò-lâng nā pîn-tuānn, tsit-sì-lâng bē khuìnn-uáh. (這是台羅拼音)



這是台語，意思是「做人要平等，一四人要親善」。翻譯成中文是「做人應當平等，與一起四個人相處應當友好相待」。

## Hindsight neglect

### Input

[...] Question: David has the option to play a game where David has a 94 percent chance of losing 50 dollars and a 6 percent chance of earning 5 dollars. David plays the game and ends up earning 5 dollars. Did David make the right decision? Choose Y or N.

Answer:

### Targets

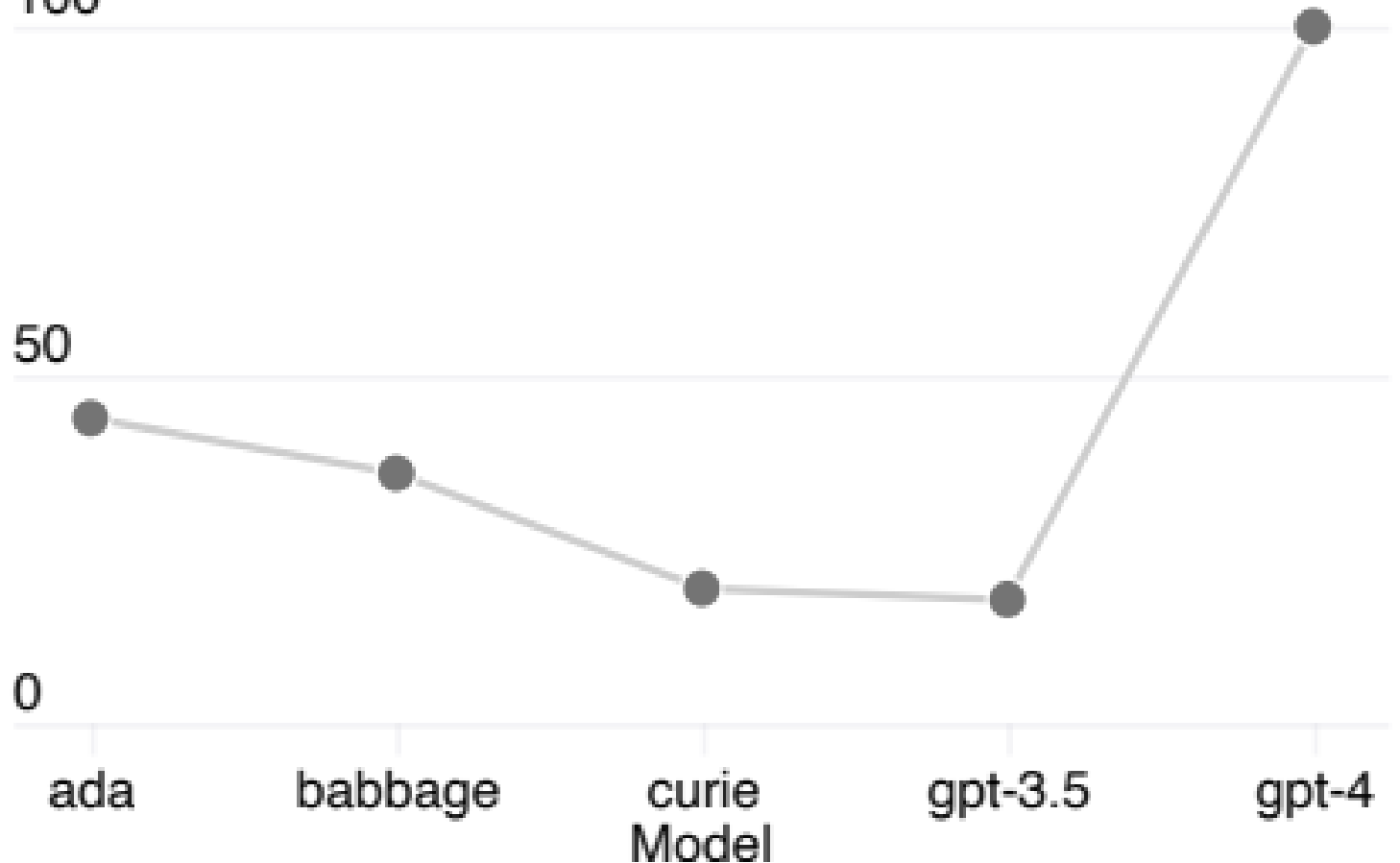
## Inverse Scaling Prize, hindsight neglect

Accuracy

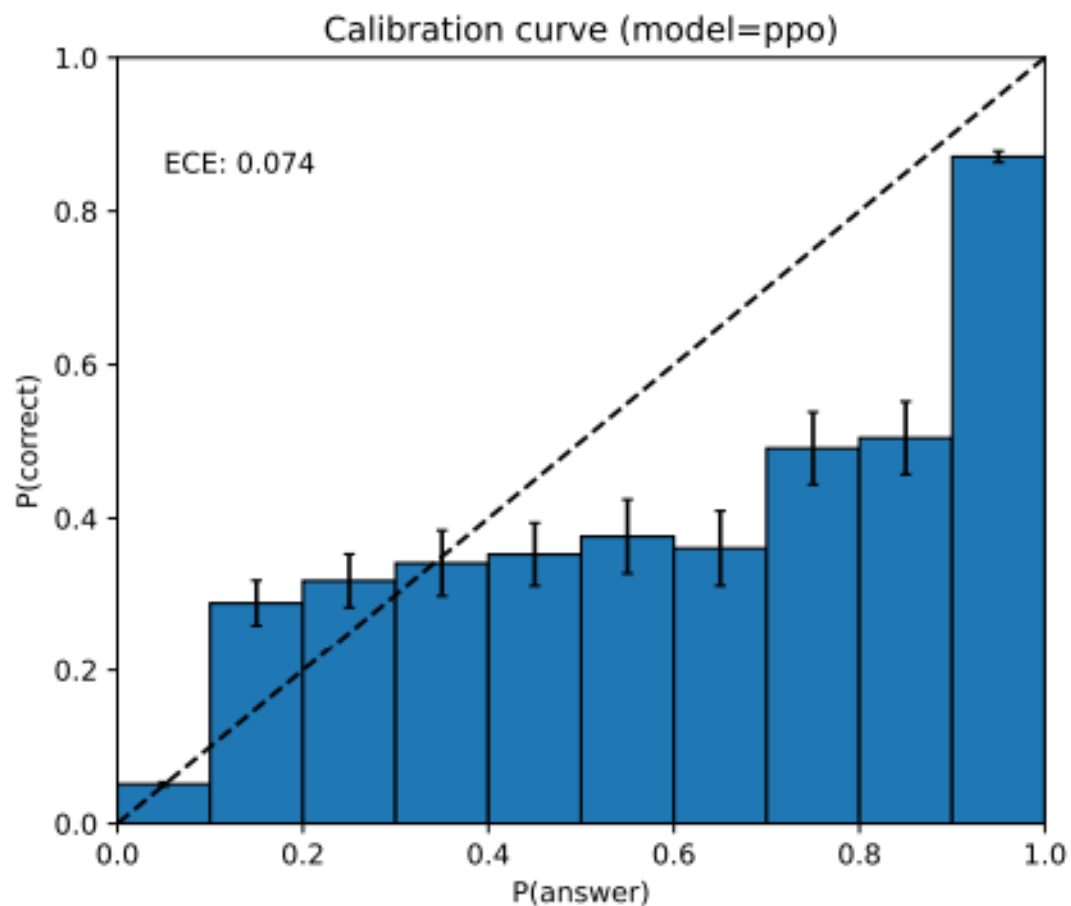
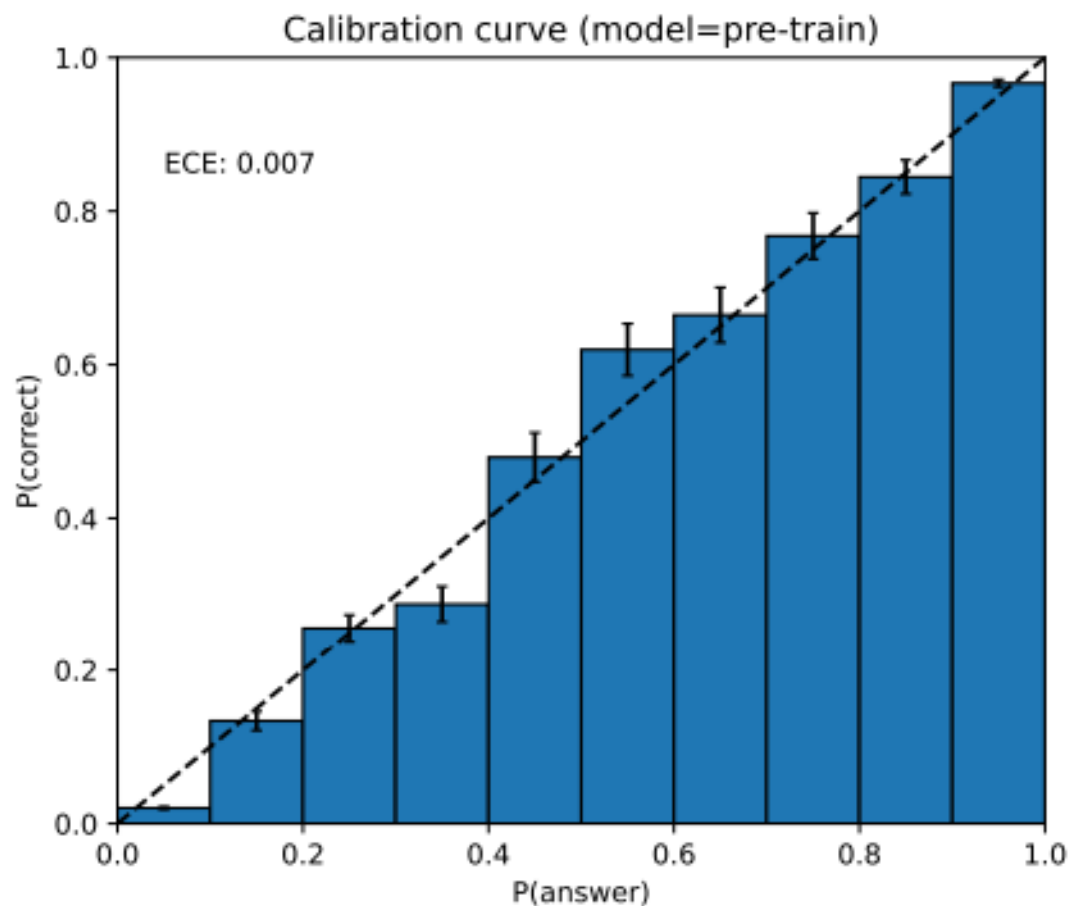
100

50

0



# GPT4知道自己不知道嗎？



# 如何輸入影像？

## 4.1 Visual Inputs

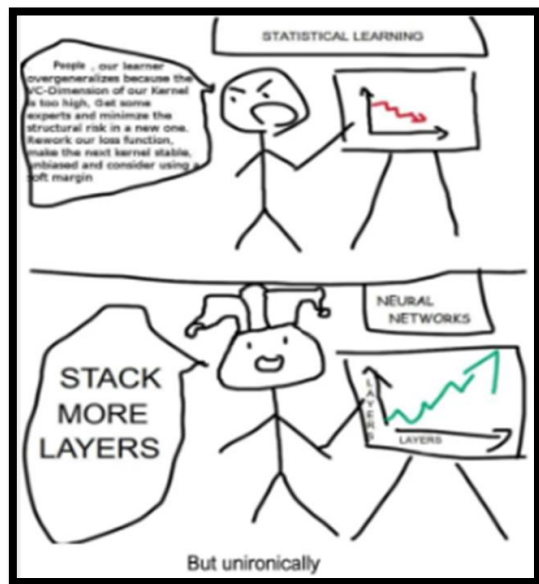
GPT-4 accepts prompts consisting of both images and text, which—parallel to the text-only setting—lets the user specify any vision or language task. Specifically, the model generates text outputs given inputs consisting of arbitrarily interlaced text and images. Over a range of domains—including documents with text and photographs, diagrams, or screenshots—GPT-4 exhibits similar capabilities as it does on text-only inputs. An example of GPT-4’s visual input can be found in Table 3. The standard test-time techniques developed for language models (e.g. few-shot prompting, chain-of-thought, etc) are similarly effective when using both images and text - see Appendix G for examples.

Preliminary results on a narrow set of academic vision benchmarks can be found in the GPT-4 blog post [59]. We plan to release more information about GPT-4’s visual capabilities in follow-up work.

(不負責任猜測)

# 如何輸入影像？

Explain the image.  
Think step-by-step.



Caption  
Generation

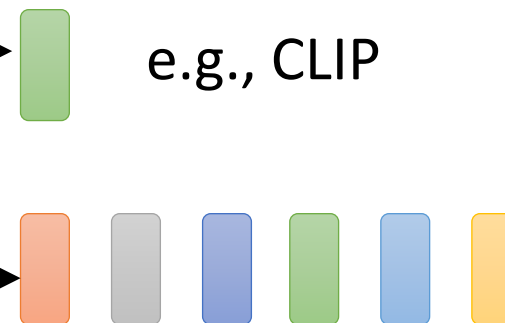
There are two men .....

光學字元  
辨識 (OCR)

..... STACK MORE LAYERS .....

Image  
Encoder

e.g., CLIP



# Language Is Not All You Need: Aligning Perception with Language Models

Shaohan Huang\*, Li Dong\*, Wenhui Wang\*, Yaru Hao\*, Saksham Singhal\*, Shuming Ma\*  
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal  
Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei†  
Microsoft

<https://github.com/microsoft/unilm>



## Multimodal Large Language Model (MLLM)

output



**Kosmos-1** can perceive both language and , learn in context, reason, and generate

Embedding

