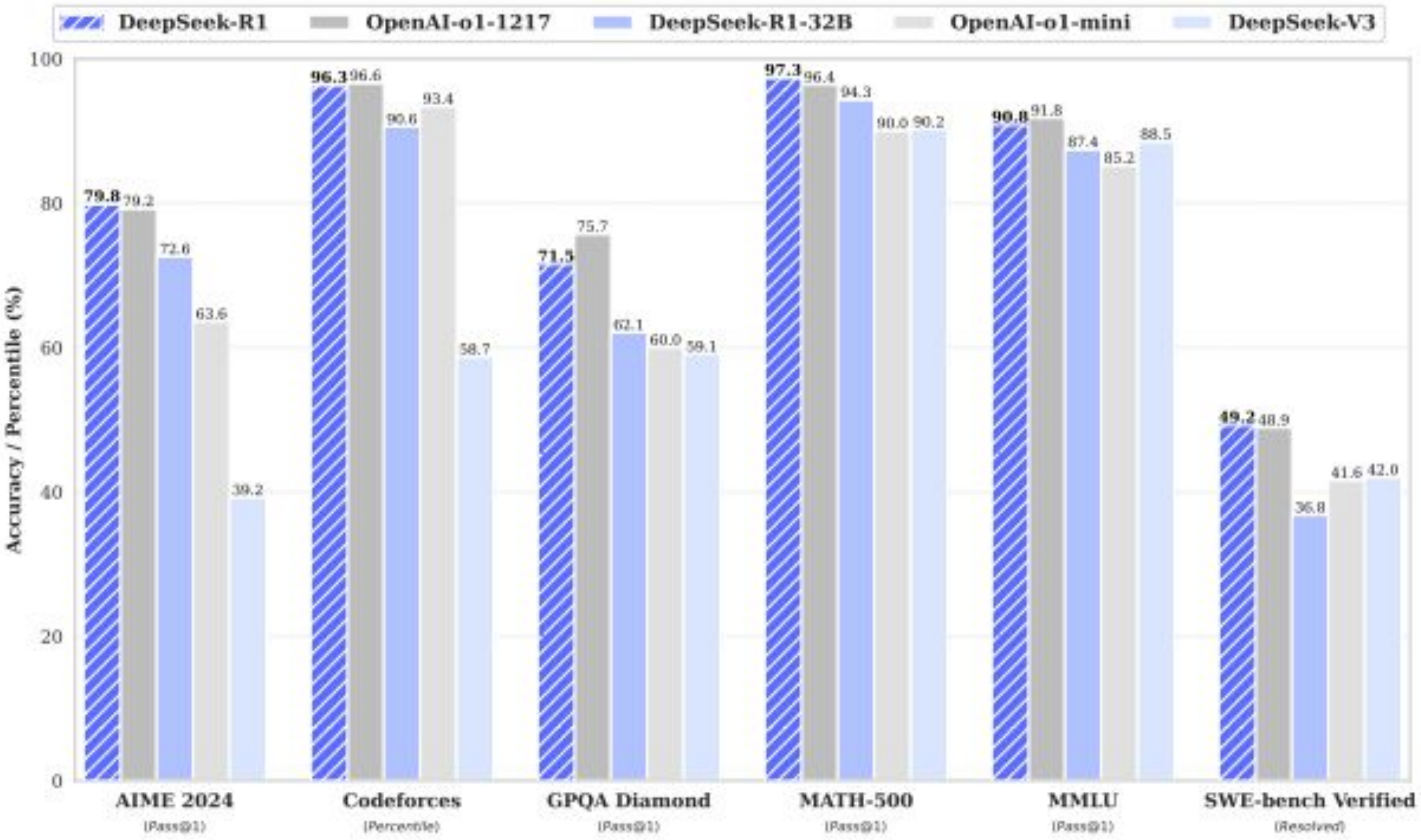
有關大型語言模型能力評量

2025/05/03

現在都用簡單粗暴的方法:解數學問題

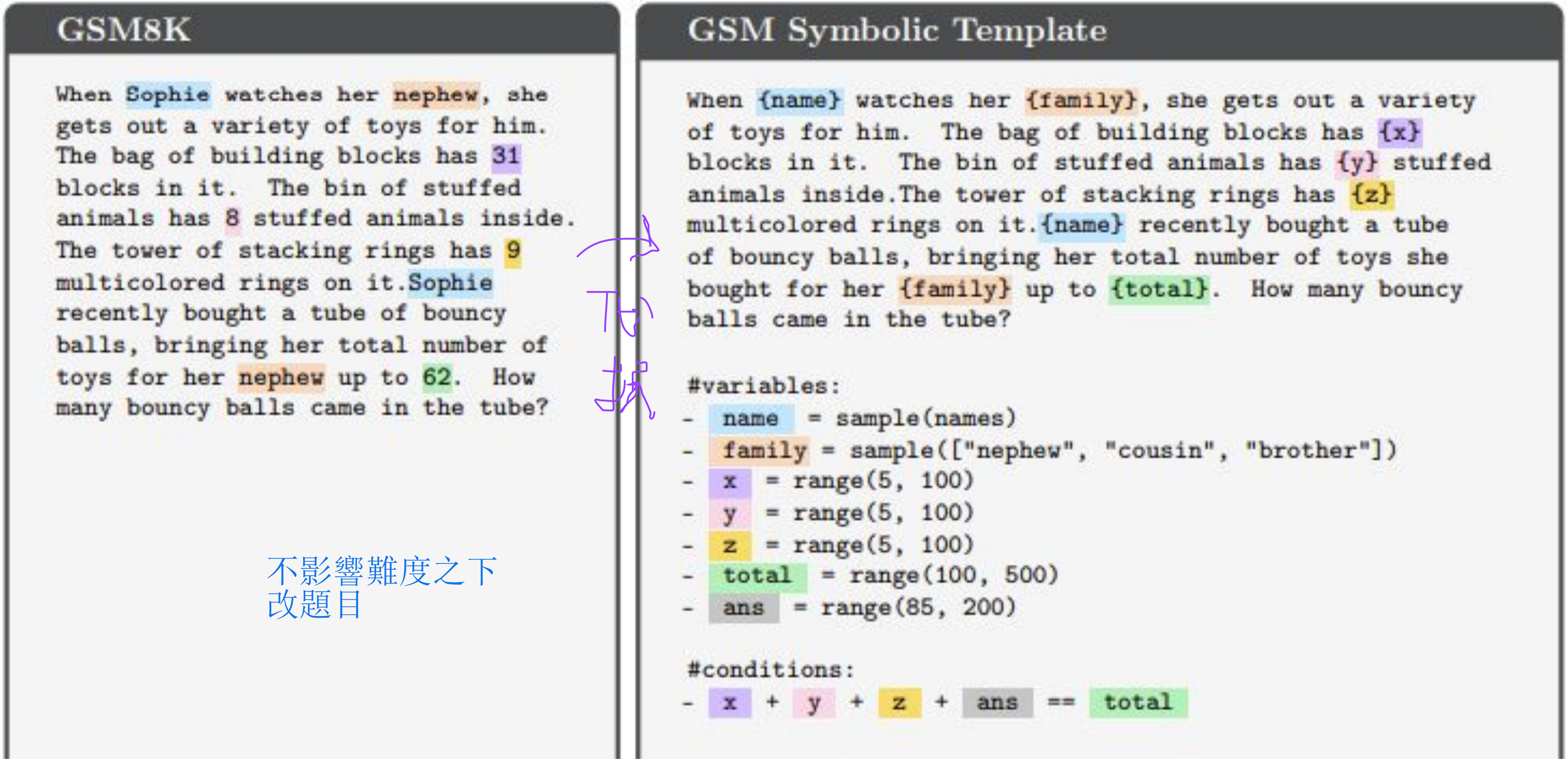
如何評量大型語言模型的「推理」能力



https://arxiv.org/abs/2501.12948

https://arxiv.org/abs/2410.05229

有多少答案可能是「記憶」出來的？



GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it.Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

代換

GSM Symbolic Template

When {name} watches her {family}, she gets out a variety of toys for him. The bag of building blocks has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside.The tower of stacking rings has {z} multicolored rings on it.{name} recently bought a tube of bouncy balls, bringing her total number of toys she bought for her {family} up to {total}. How many bouncy balls came in the tube?

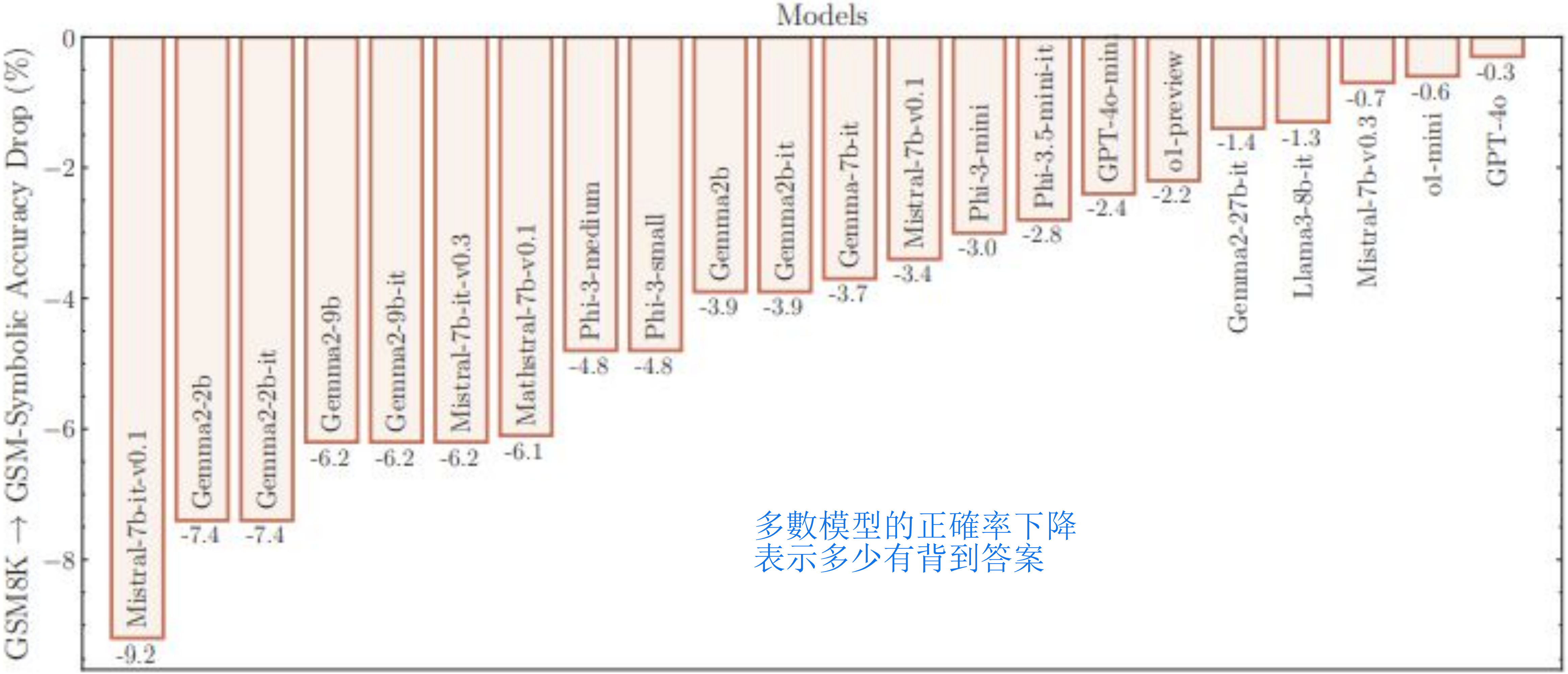
```
#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x   = range(5, 100)
- y   = range(5, 100)
- z   = range(5, 100)
- total  = range(100, 500)
- ans   = range(85, 200)

#conditions:
- x + y + z + ans == total
```

不影響難度之下
改題目

https://arxiv.org/abs/2410.05229

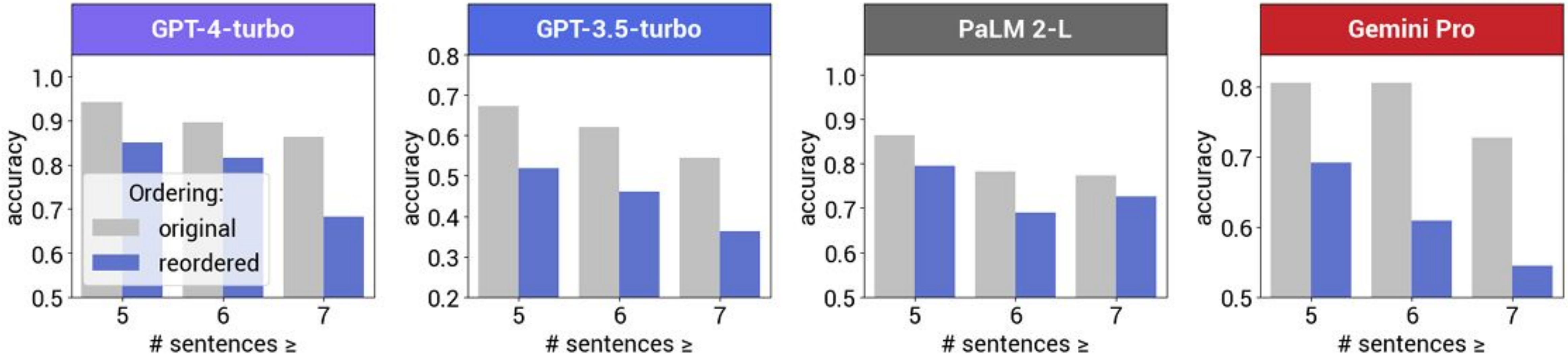
有多少答案可能是「記憶」出來的？



多數模型的正確率下降
表示多少有背到答案

有多少答案可能是「記憶」出來的？

在不影響題意的情況下，把句子順序換掉，正確率下降

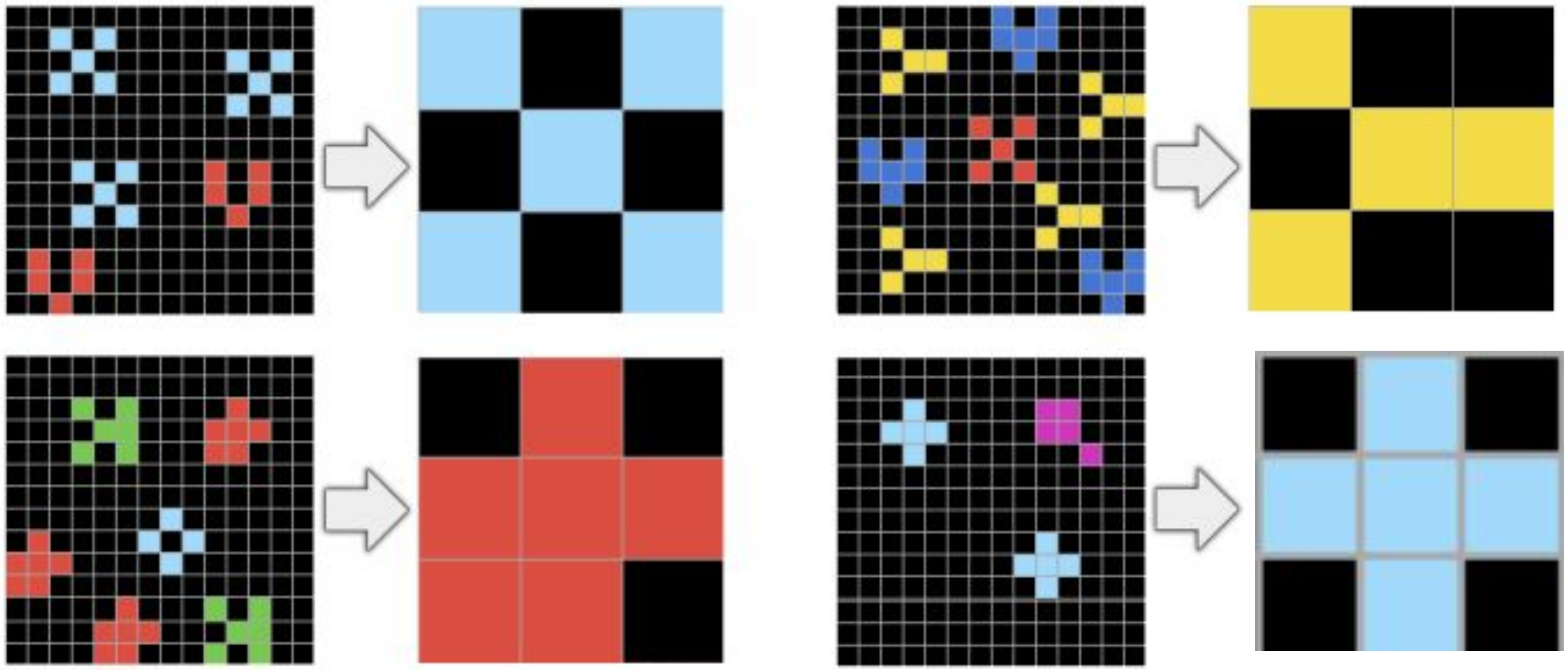


https://arxiv.org/abs/2402.08939

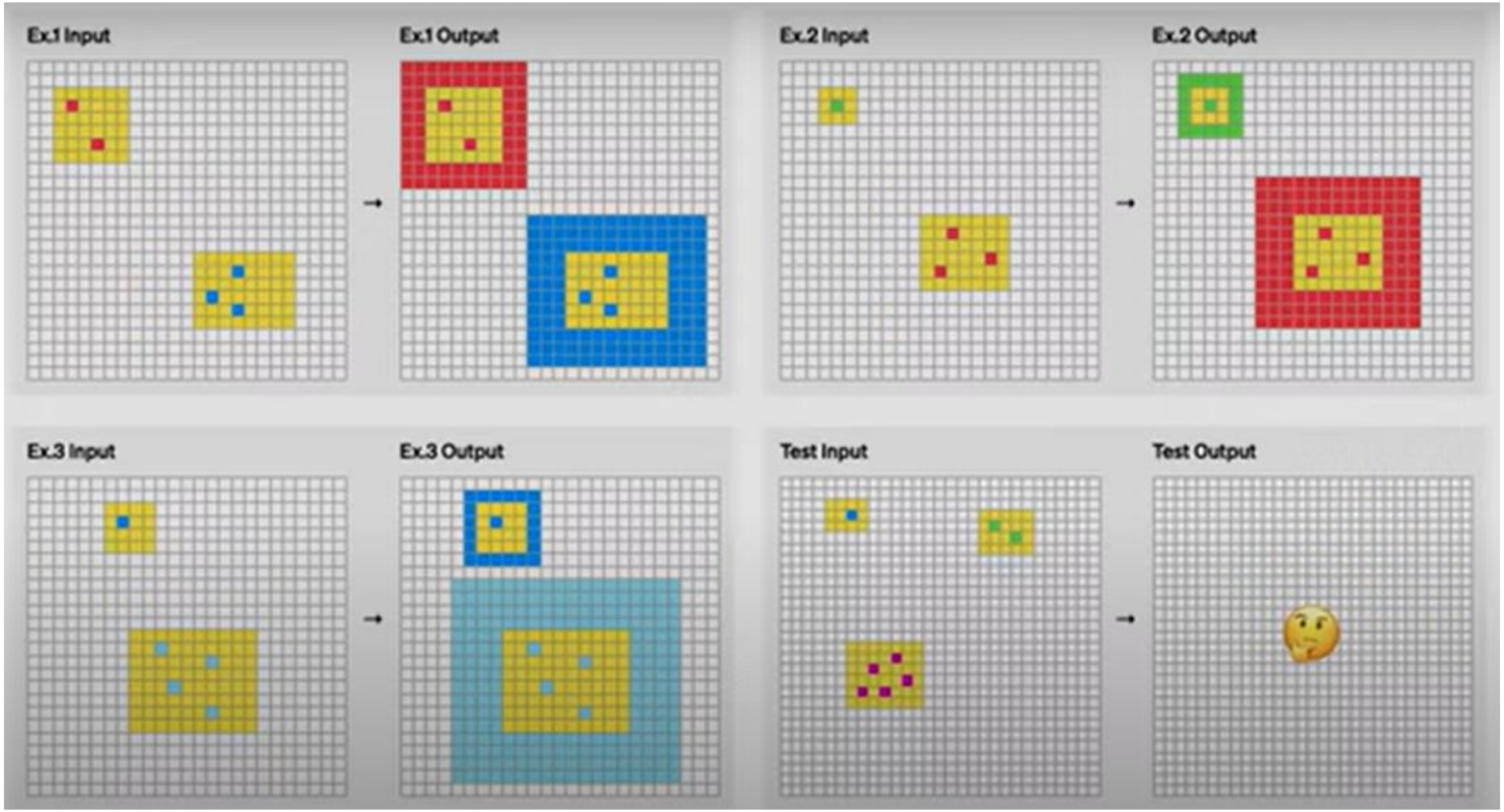
現在熱門的模型推理能力測試

Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI)

https://arxiv.org/abs/1911.01547



ARC-AGI



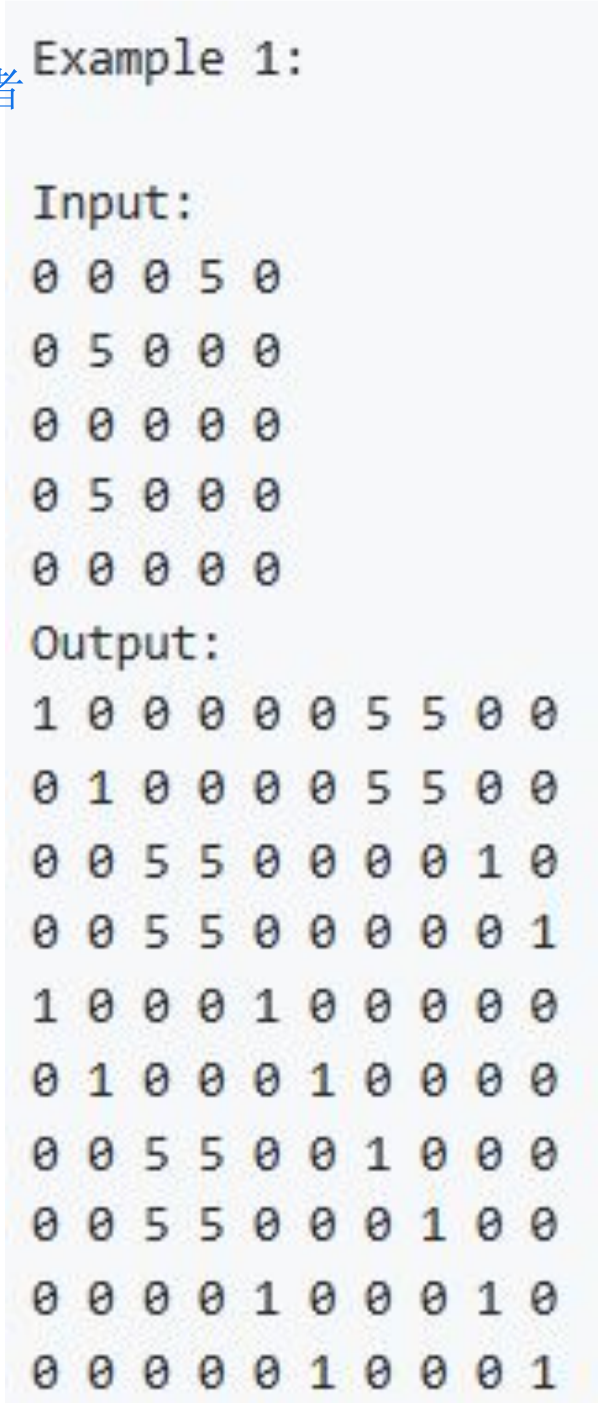
https://www.youtube.com/watch?v=SKBG1sqdyIU

同時也是keras的作者

ARC-AGI

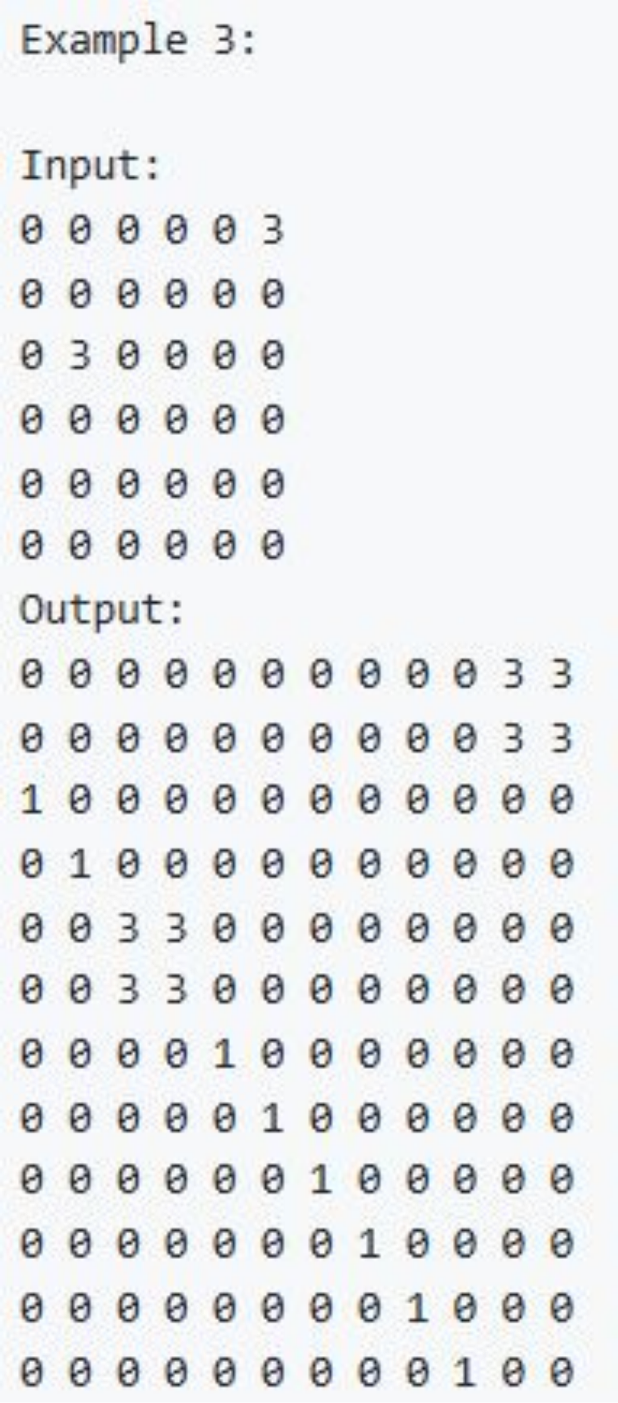
1-5表示不同顏色
0表示沒有顏色

https://github.com/arcprize/model_baseline/blob/main/prompt_example_o3.md



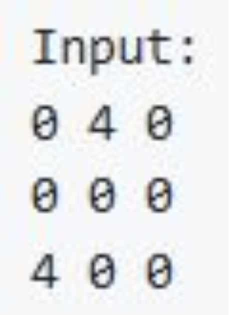
```
Example 1:

Input:
0 0 0 5 0
0 5 0 0 0
0 0 0 0 0
0 5 0 0 0
0 0 0 0 0
Output:
1 0 0 0 0 0 5 5 0 0
0 1 0 0 0 0 5 5 0 0
0 0 5 5 0 0 0 0 1 0
0 0 5 5 0 0 0 0 0 1
1 0 0 0 1 0 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 0 5 5 0 0 1 0 0 0
0 0 5 5 0 0 0 1 0 0
0 0 0 0 1 0 0 0 1 0
0 0 0 0 0 1 0 0 0 1
```

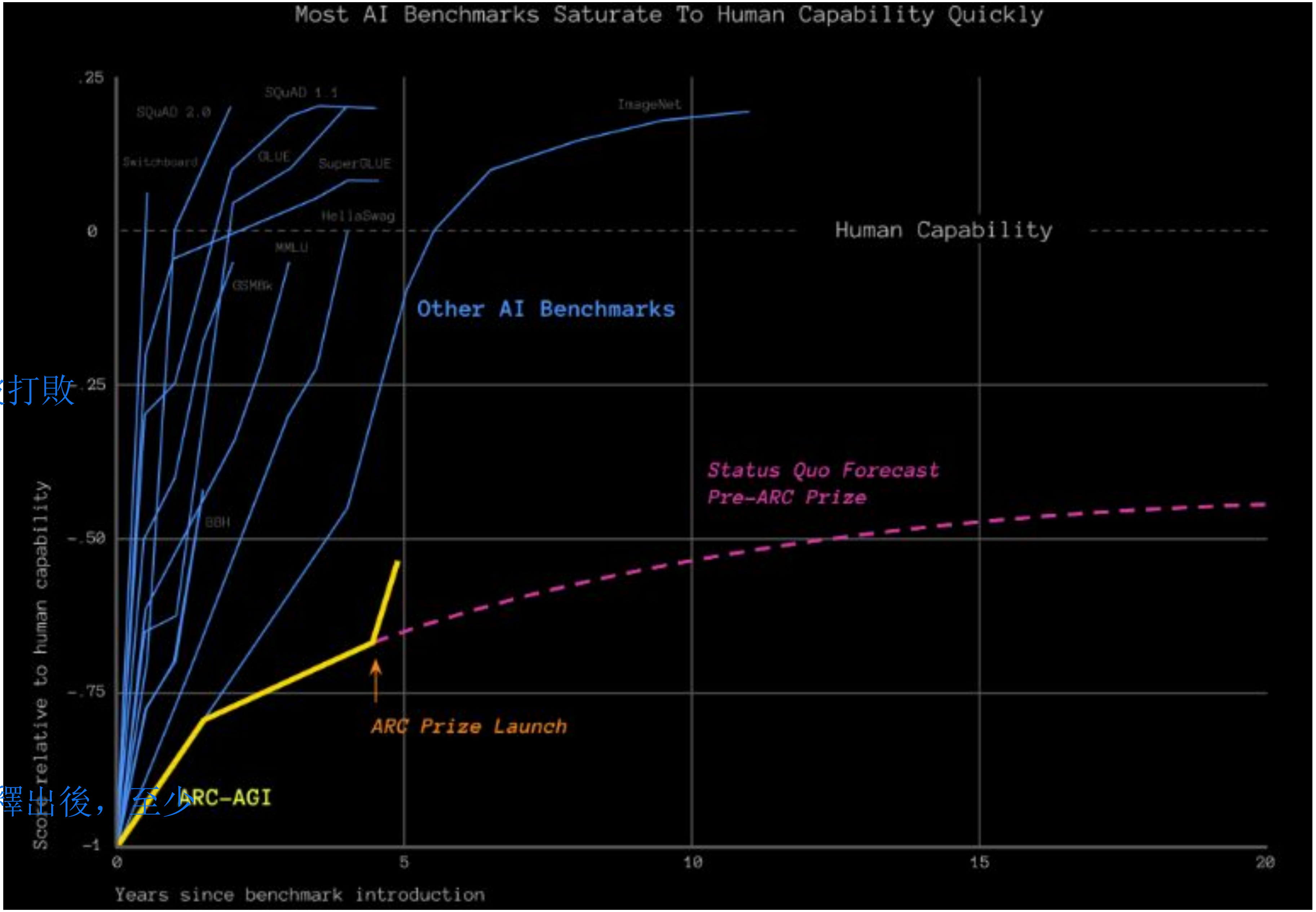


```
Example 3:

Input:
0 0 0 0 0 3
0 0 0 0 0 0
0 3 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
Output:
0 0 0 0 0 0 0 0 0 0 3 3
0 0 0 0 0 0 0 0 0 0 3 3
1 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 0 0 0
0 0 3 3 0 0 0 0 0 0 0 0
0 0 3 3 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0
```



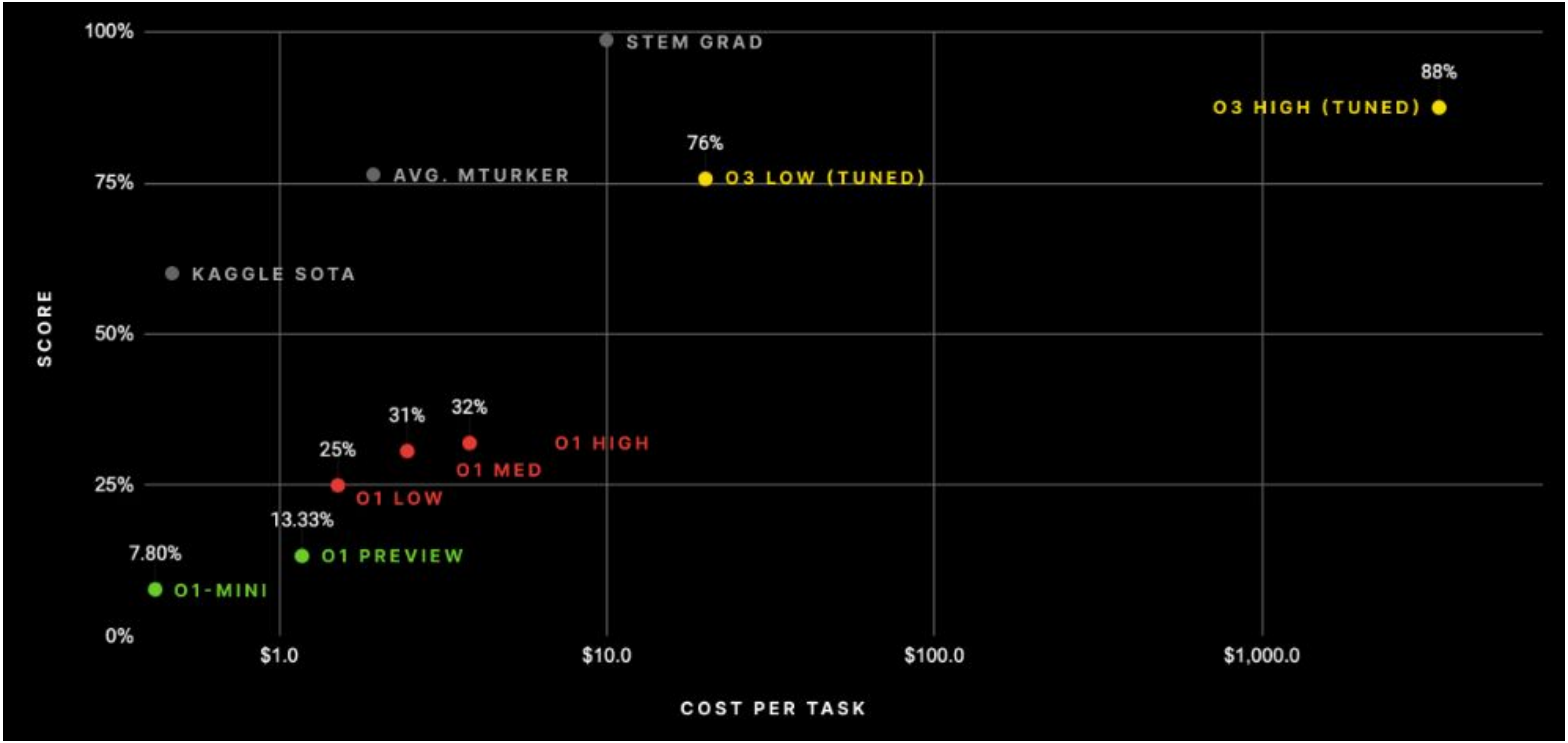
```
Input:
0 4 0
0 0 0
4 0 0
```



多數corpus很多被打敗

arc-agi 2019年釋出後，至少
5年還沒被破解
直到o3模型

https://arcprize.org/blog/3-month-update o3比一般學生強，弱於stem學生

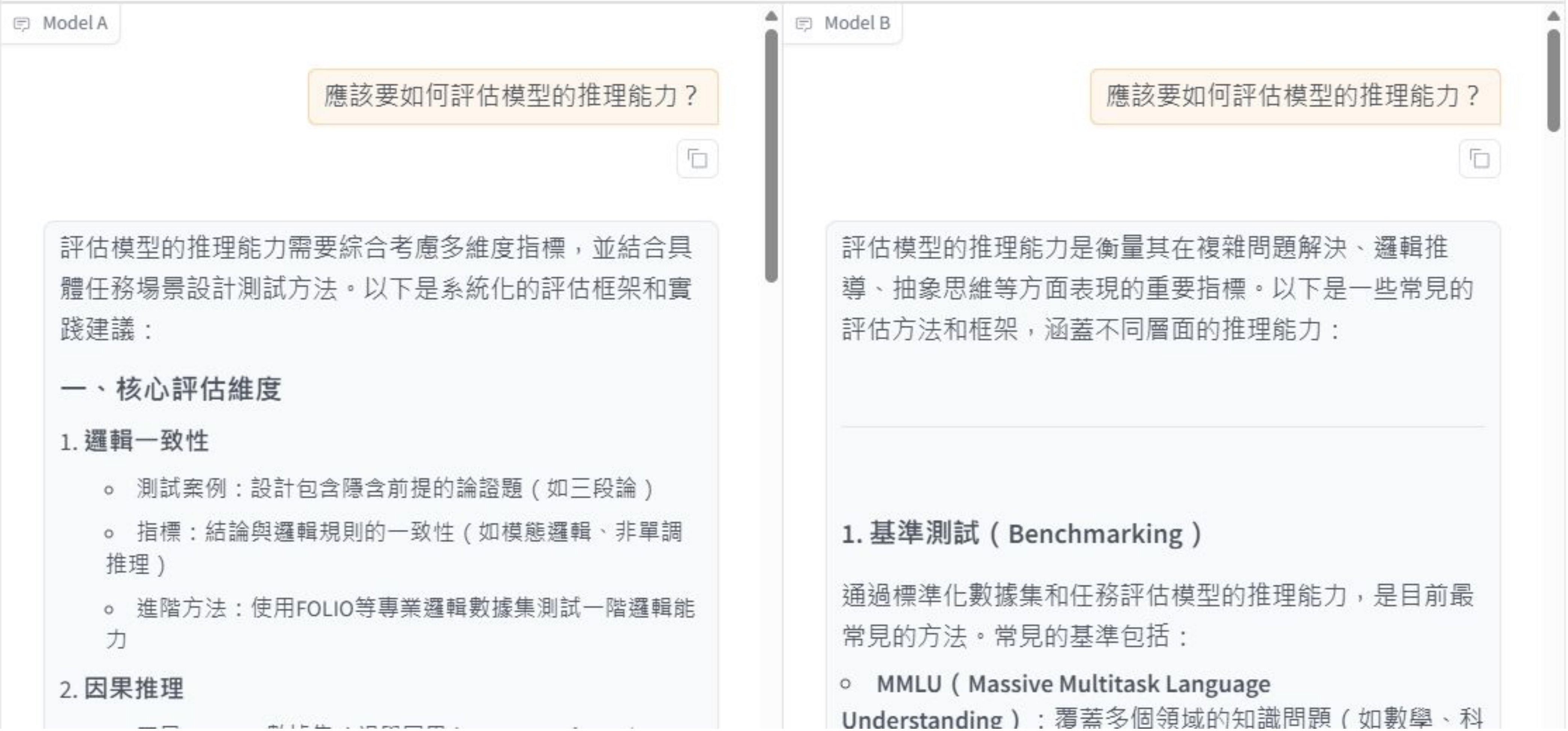


只要有固定出題方向就可能被猜題->有人想到chatbot arena

https://arcprize.org/blog/oai-o3-pub-breakthrough

Chatbot Arena

但實際上人類還是有偏好方向
ex.較多的emoji,比較長的輸出等



應該要如何評估模型的推理能力？

評估模型的推理能力需要綜合考慮多維度指標，並結合具體任務場景設計測試方法。以下是系統化的評估框架和實踐建議：

一、核心評估維度

1. **邏輯一致性**

- 測試案例：設計包含隱含前提的論證題（如三段論）
- 指標：結論與邏輯規則的一致性（如模態邏輯、非單調推理）
- 進階方法：使用FOLIO等專業邏輯數據集測試一階邏輯能力

2. **因果推理**

應該要如何評估模型的推理能力？

評估模型的推理能力是衡量其在複雜問題解決、邏輯推導、抽象思維等方面表現的重要指標。以下是一些常見的評估方法和框架，涵蓋不同層面的推理能力：

---

1. 基準測試（Benchmarking）

通過標準化數據集和任務評估模型的推理能力，是目前最常見的方法。常見的基準包括：

- **MMLU（Massive Multitask Language Understanding）**：覆蓋多個領域的知識問題（如數學、科

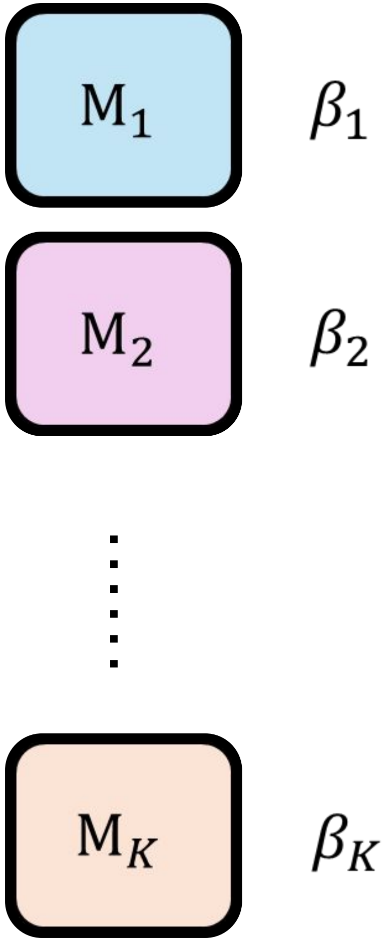
https://lmarena.ai/

官方試圖解決問題：

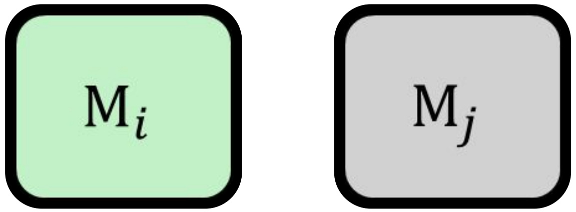
https://blog.lmarena.ai/blog/2024/style-control/

Chatbot Arena - Elo Score

模型



隨機匹配到的模型



$$\frac{1}{1+exp\left(-\frac{\beta_i-\beta_j}{400}\right)} = E_{i,j}$$

根據比賽結果統計勝率

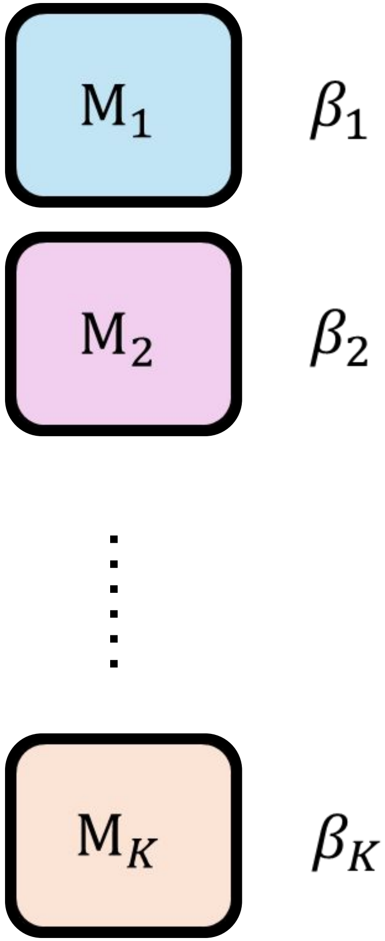
算出 $\beta_1, \beta_2, \dots, \beta_K$

第i個模型戰力-第j個模型的戰力
除掉normalization的分數
為了讓分數好看一點通常會設定400
乘上負號再取exponencial
就是sigmoid function

i的戰力若>>>j的話，算出來會趨近1
i如果<<<j，減出來的就會是負值，算出來會趨近0

https://blog.lmarena.ai/blog/2024/style-control/

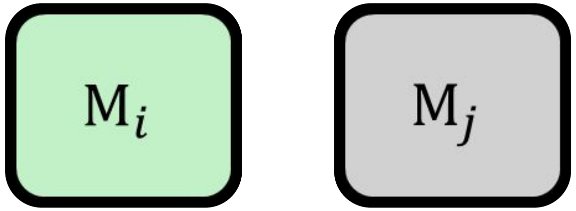
Chatbot Arena - Elo Score


M_1 β_1

M_2 β_2

$\vdots$

M_K β_K


M_i M_j

$$\frac{1}{1+exp\left(-\frac{\beta_i-\beta_j+\beta_0}{400}\right)} = E_{i,j}$$

根據比賽結果統計勝率

算出 $\beta_1, \beta_2, \dots, \beta_K$

算出 $\gamma_1, \gamma_2, \dots$

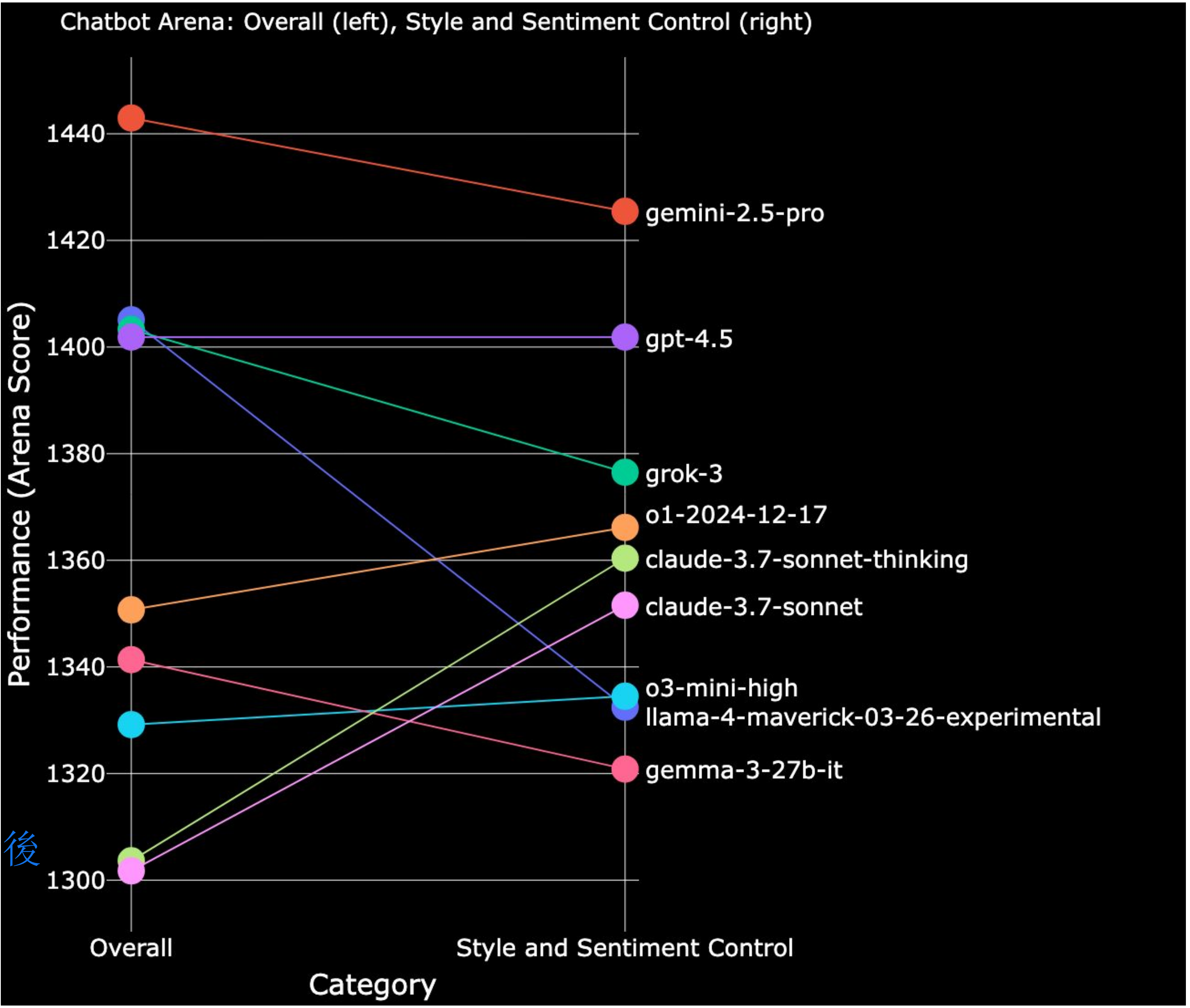
類似棋類遊戲會考慮先手優勢

β_0 模型實力以外的因素

$$\beta_0 = \gamma_1(\text{答案長度差}) + \gamma_2(emoji\ \text{數量差}) + \cdots$$

人類偏好

有無考慮與
風格相關的因素
BO，會影響模型排名



EX.
Claude因為比較少
輸出emoji等，比較
不討喜，但去除此因素後
排名往前

就算是chatbot arena也可能被hack

https://blog.lmarena.ai/blog/2025/sentiment-control/

Goodhart’s law

- 一項指標一旦被當作目標，它就不再是一個好的指標。

給錢抓蛇，結果民眾反而養一堆蛇