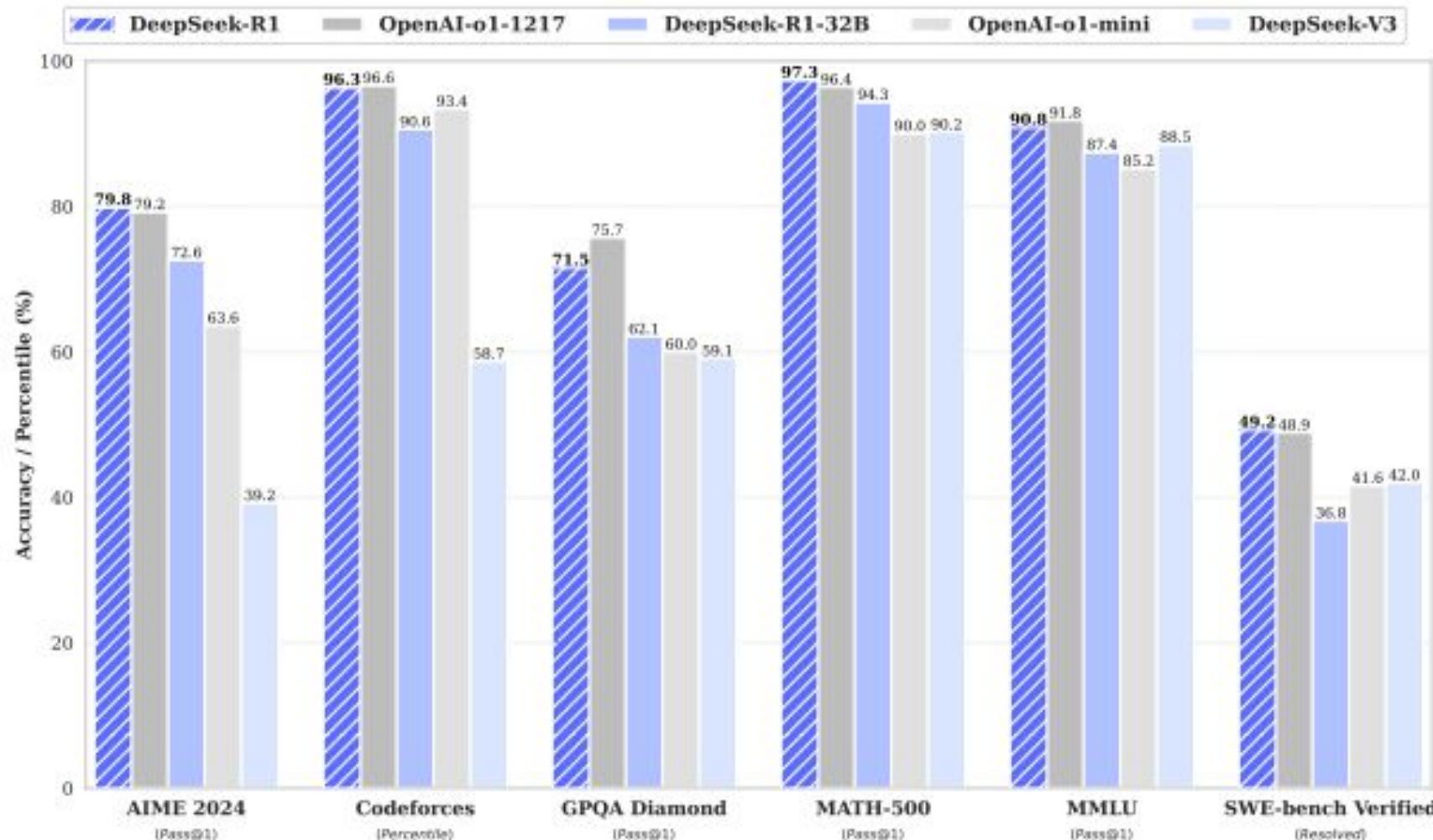


有關大型語言模型 能力評量

2025/05/03

如何評量大型語言模型的「推理」能力



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?

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?

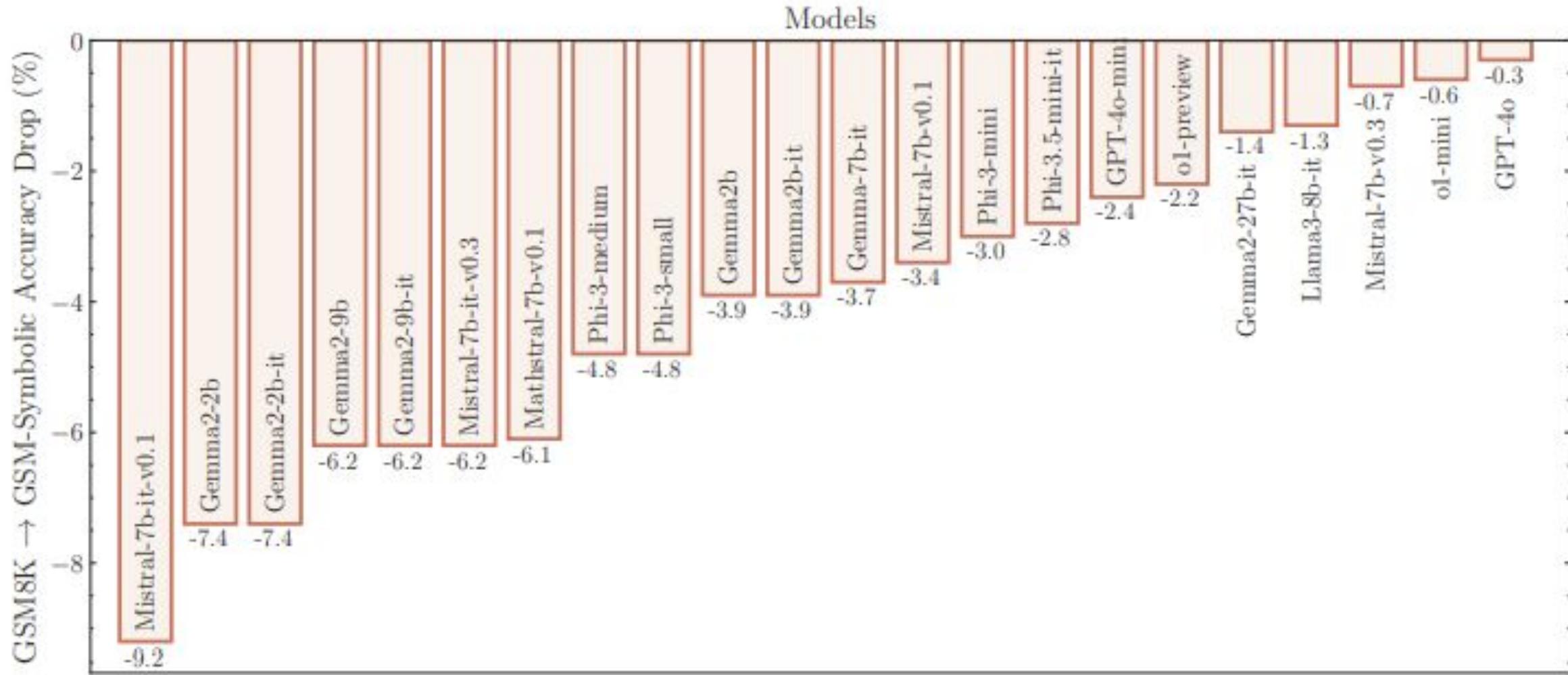
#variables:

- 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)

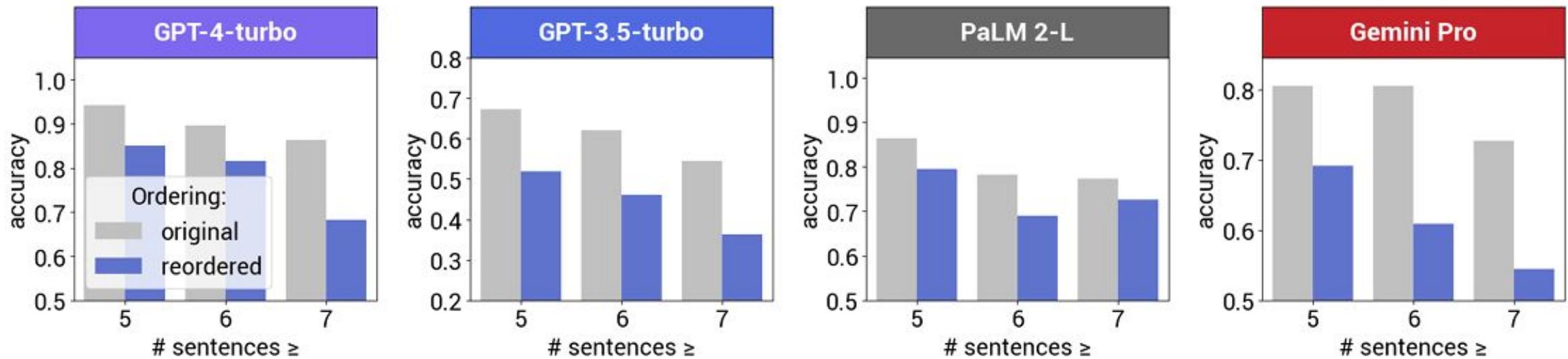
#conditions:

- x + y + z + ans == total

有多少答案可能是「記憶」出來的？



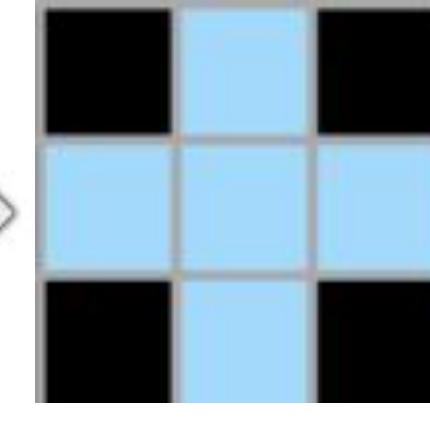
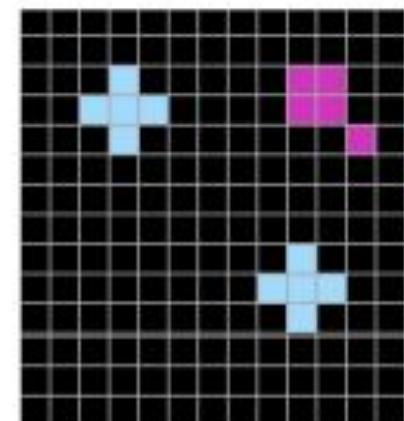
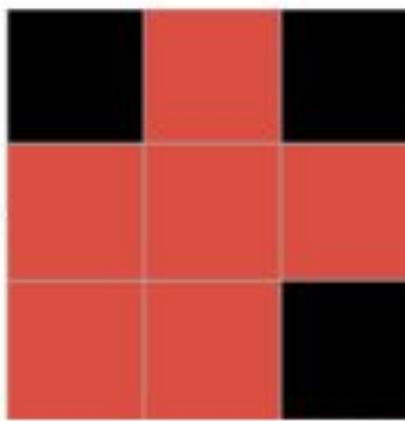
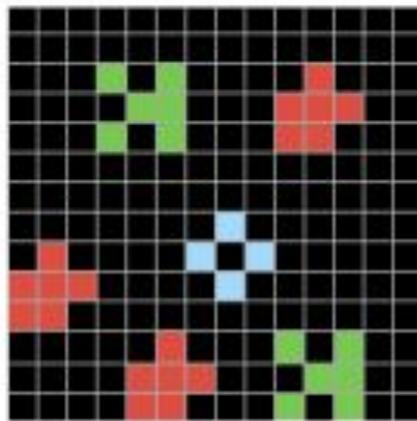
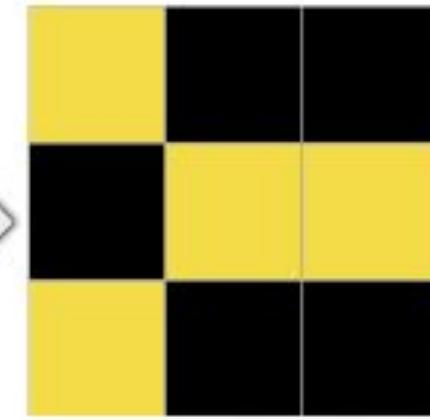
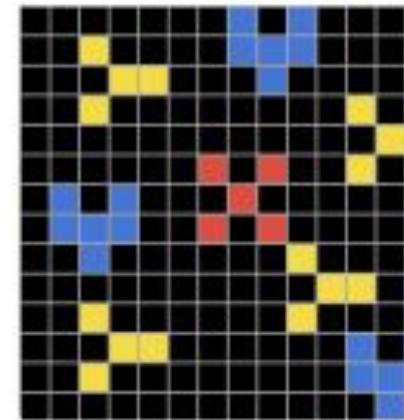
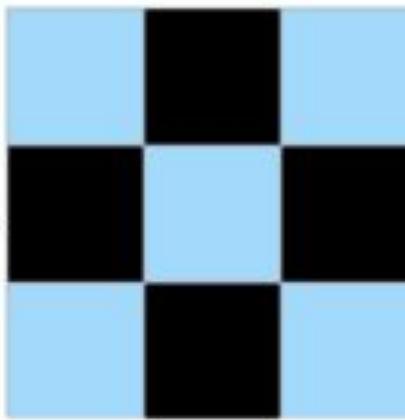
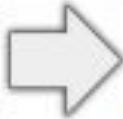
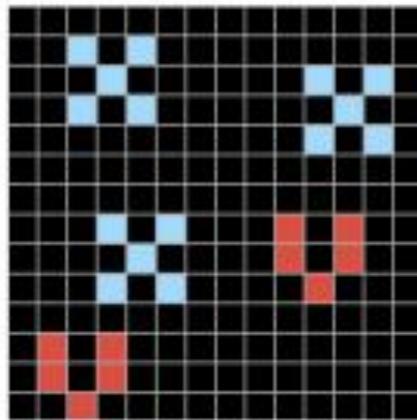
有多少答案可能是「記憶」出來的？



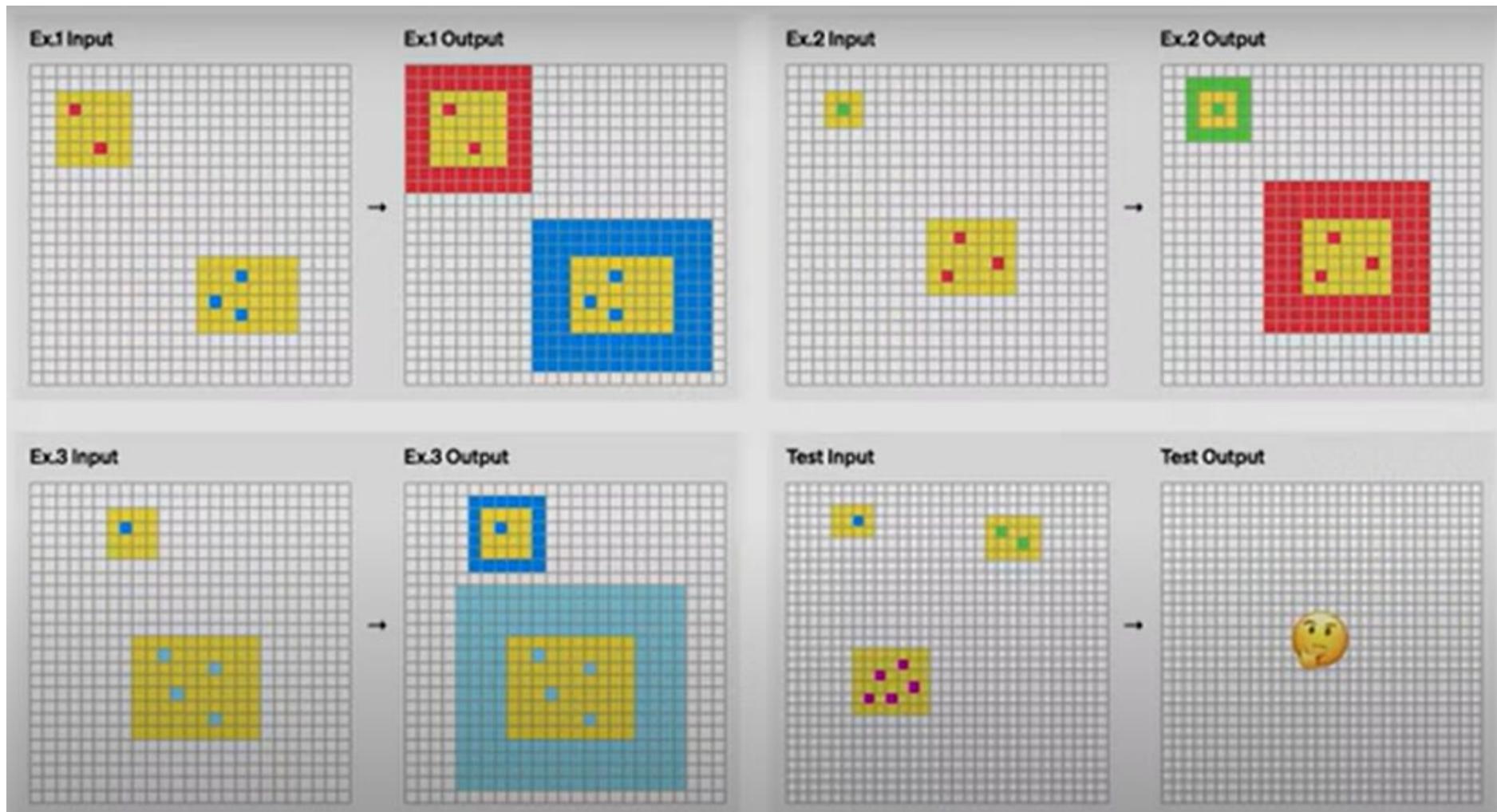
<https://arxiv.org/abs/2402.08939>

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>

ARC-AGI

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

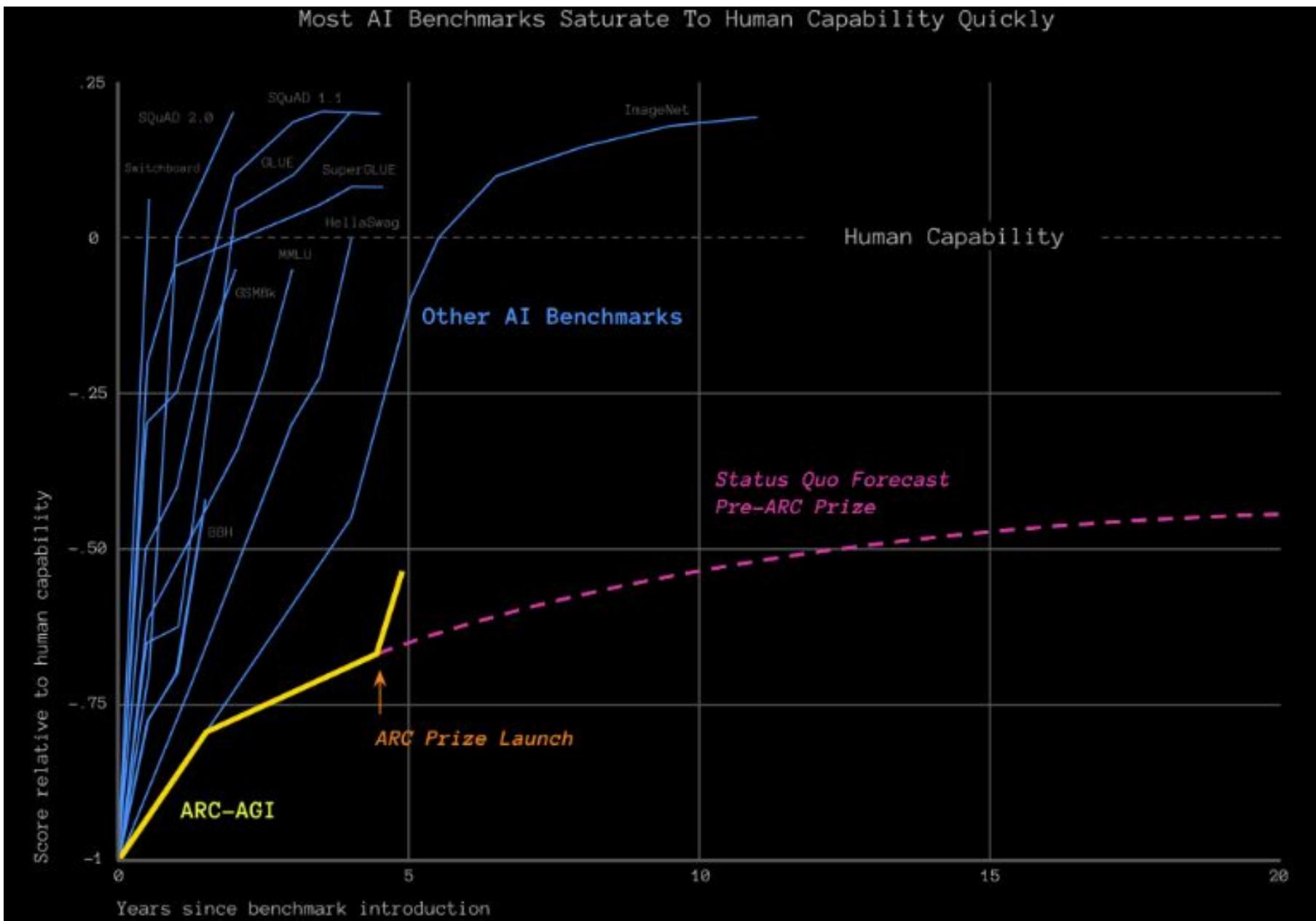
Input:

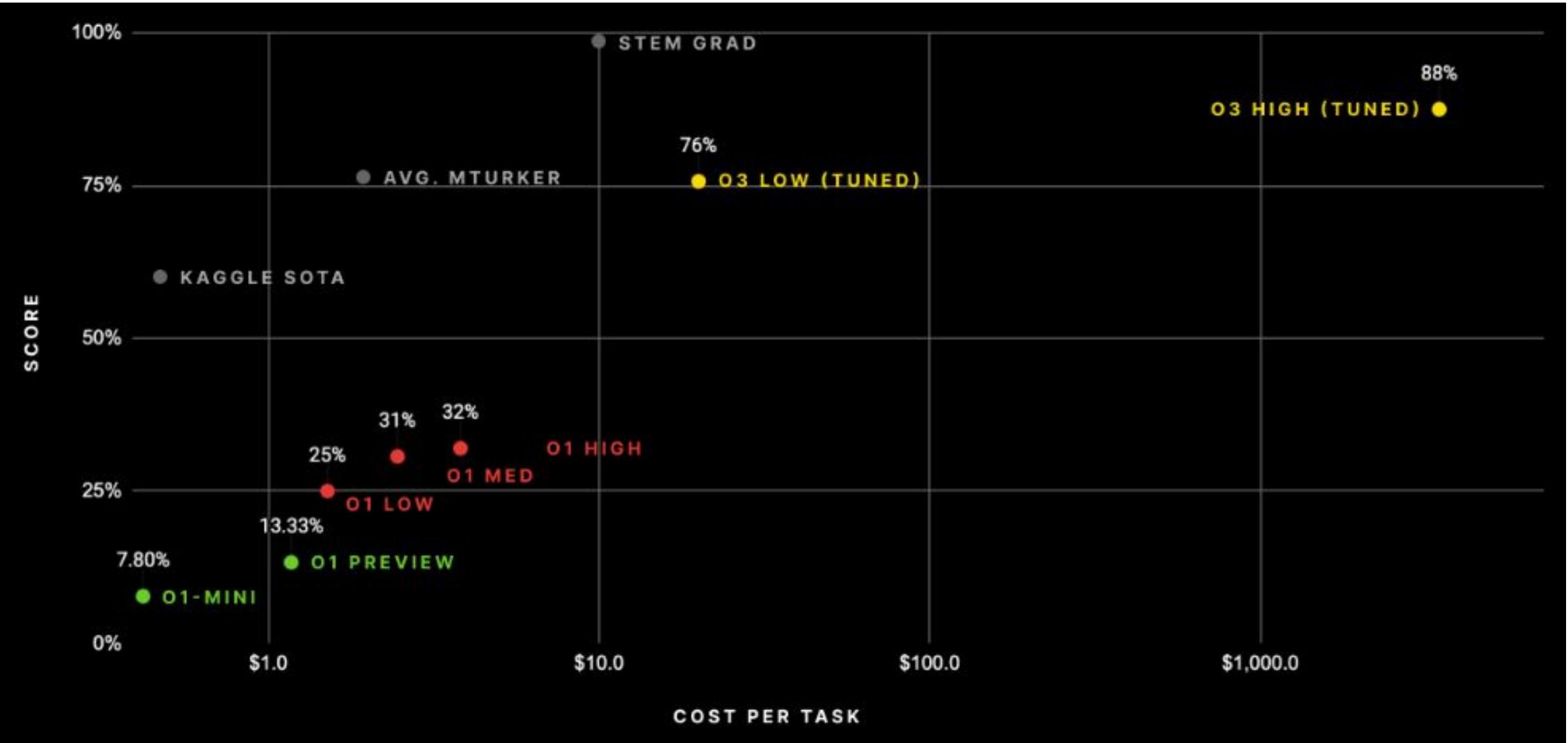
0 4 0

0 0 0

4 0 0

https://github.com/arcpriz/e/model_baseline/blob/main/prompt_example_o3.md





Chatbot Arena

The diagram illustrates a comparison between two models, Model A and Model B, focusing on how to evaluate their reasoning abilities. Both models are shown with a question box asking '應該要如何評估模型的推理性？' (How should we evaluate a model's reasoning ability?) and a response box below it.

Model A: The response box contains text about evaluating reasoning abilities across various dimensions and provides practical advice. It also includes a section on core evaluation dimensions and specific methods for logic consistency and causal reasoning.

Model B: The response box contains text about evaluating reasoning abilities across various dimensions and provides practical advice. It also includes a section on benchmarking and mentions MMLU (Massive Multitask Language Understanding) as a common method.

评估模型的推理性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**

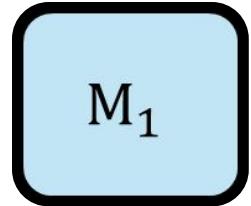
評估模型的推理性是衡量其在複雜問題解決、邏輯推導、抽象思維等方面表現的重要指標。以下是一些常見的評估方法和框架，涵蓋不同層面的推理性：

- 1. 基準測試 (Benchmarking)**

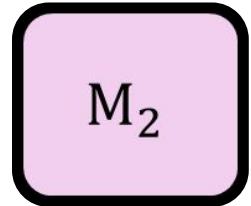
通過標準化數據集和任務評估模型的推理性，是最常見的方法。常見的基準包括：

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

Chatbot Arena - Elo Score



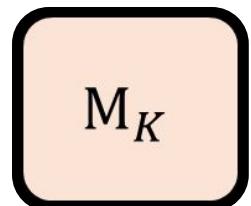
β_1



M_2

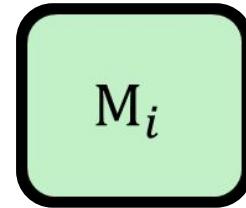
β_2

⋮

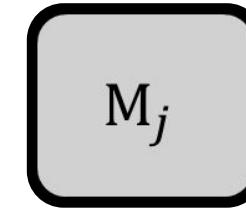


M_K

β_K



M_i

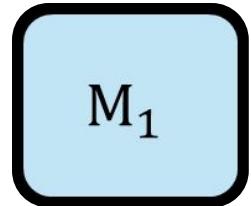


M_j

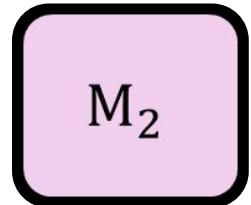
$$\frac{1}{1 + \exp\left(-\frac{\beta_i - \beta_j}{400}\right)} = E_{i,j}$$

根據比賽結果統計勝率
算出 $\beta_1, \beta_2, \dots, \beta_K$

Chatbot Arena - Elo Score

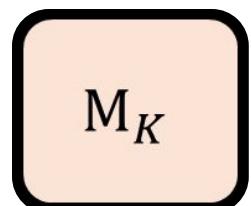


β_1

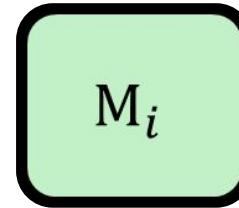


β_2

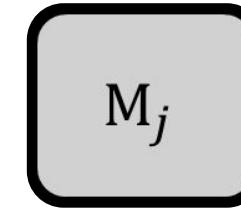
⋮



β_K



M_i



M_j

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

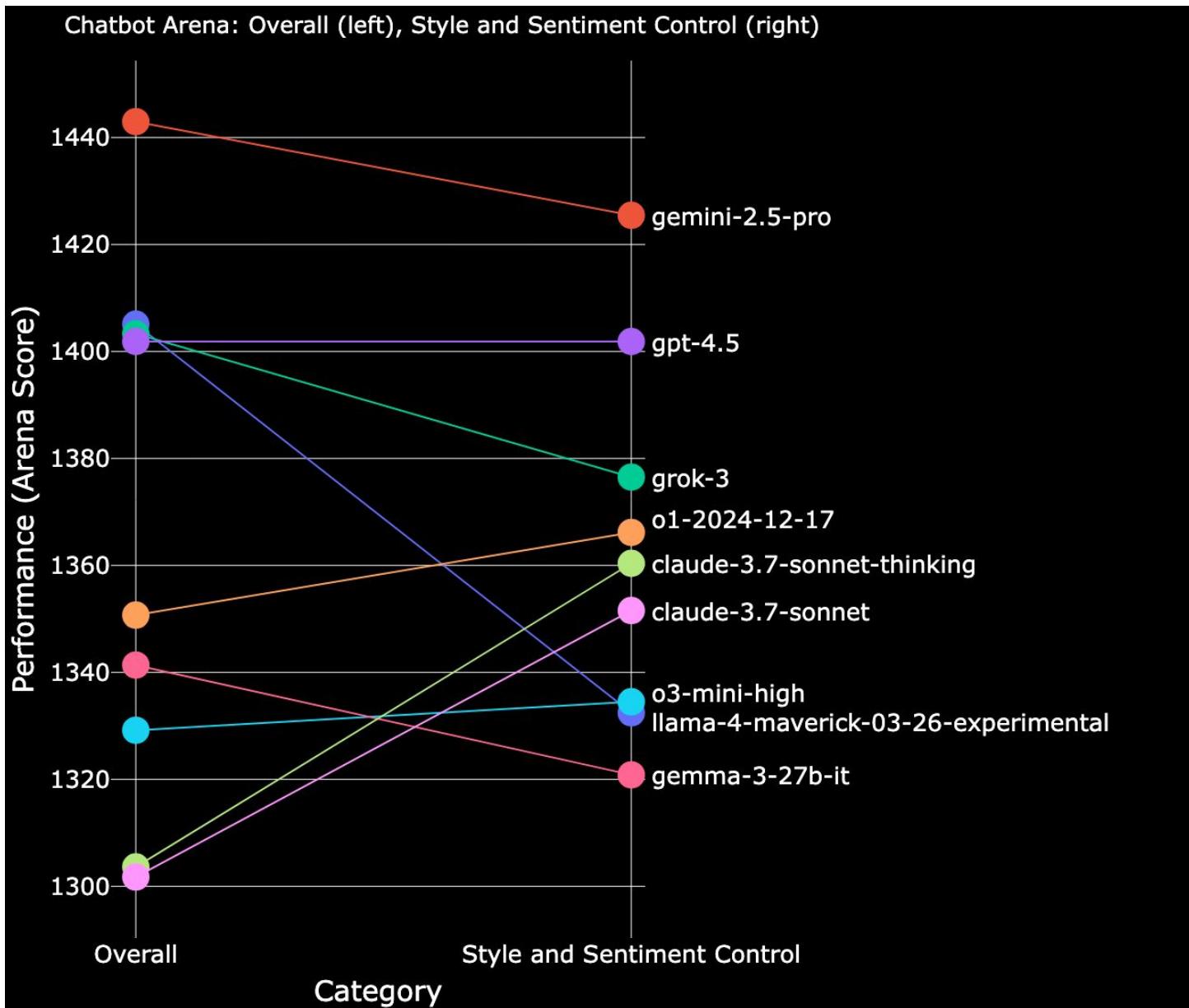
根據比賽結果統計勝率

算出 $\beta_1, \beta_2, \dots, \beta_K$

β_0 模型實力以外的因素

算出 $\gamma_1, \gamma_2, \dots$

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



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。

