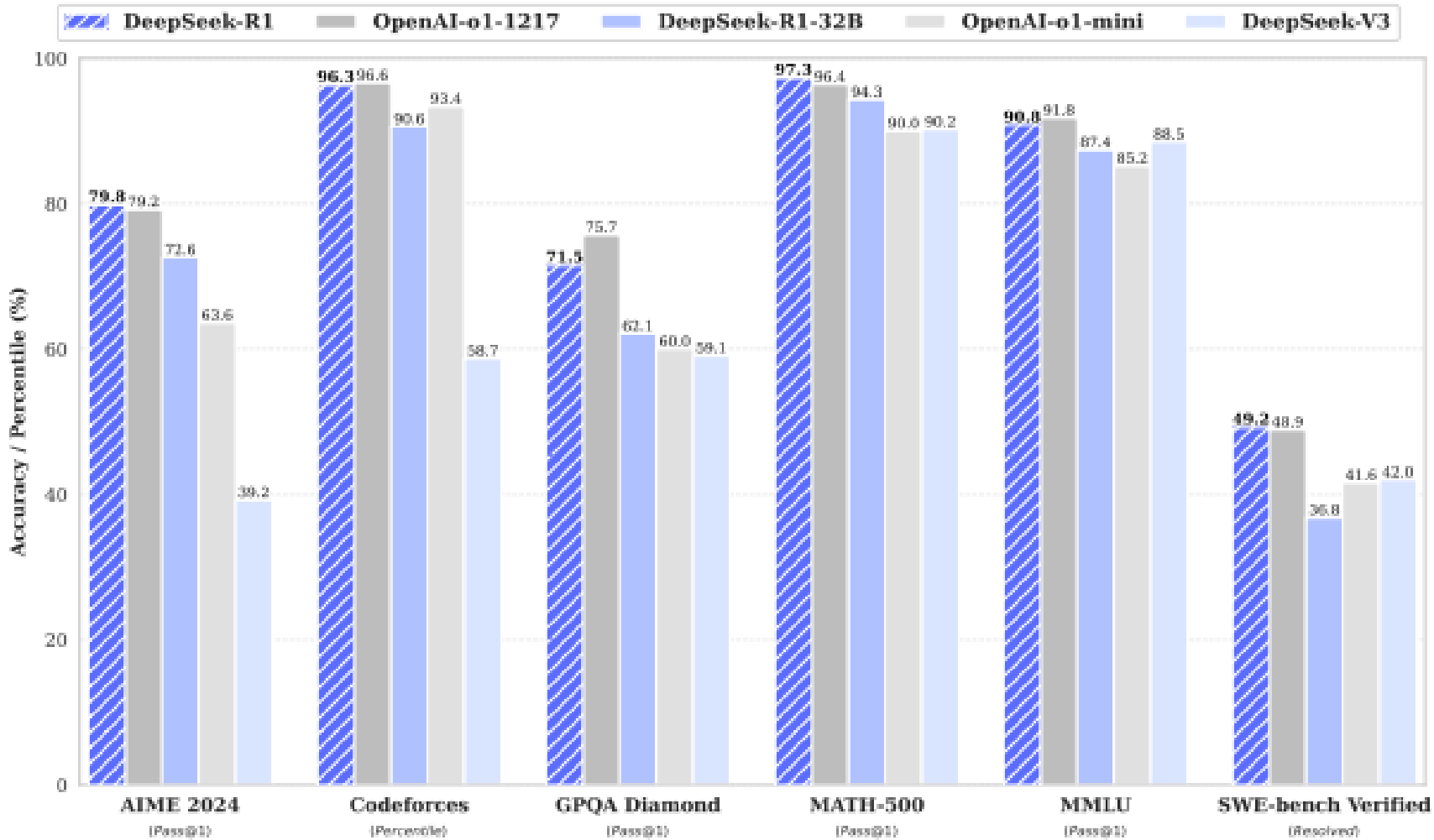
有關大型語言模型能力評量

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?

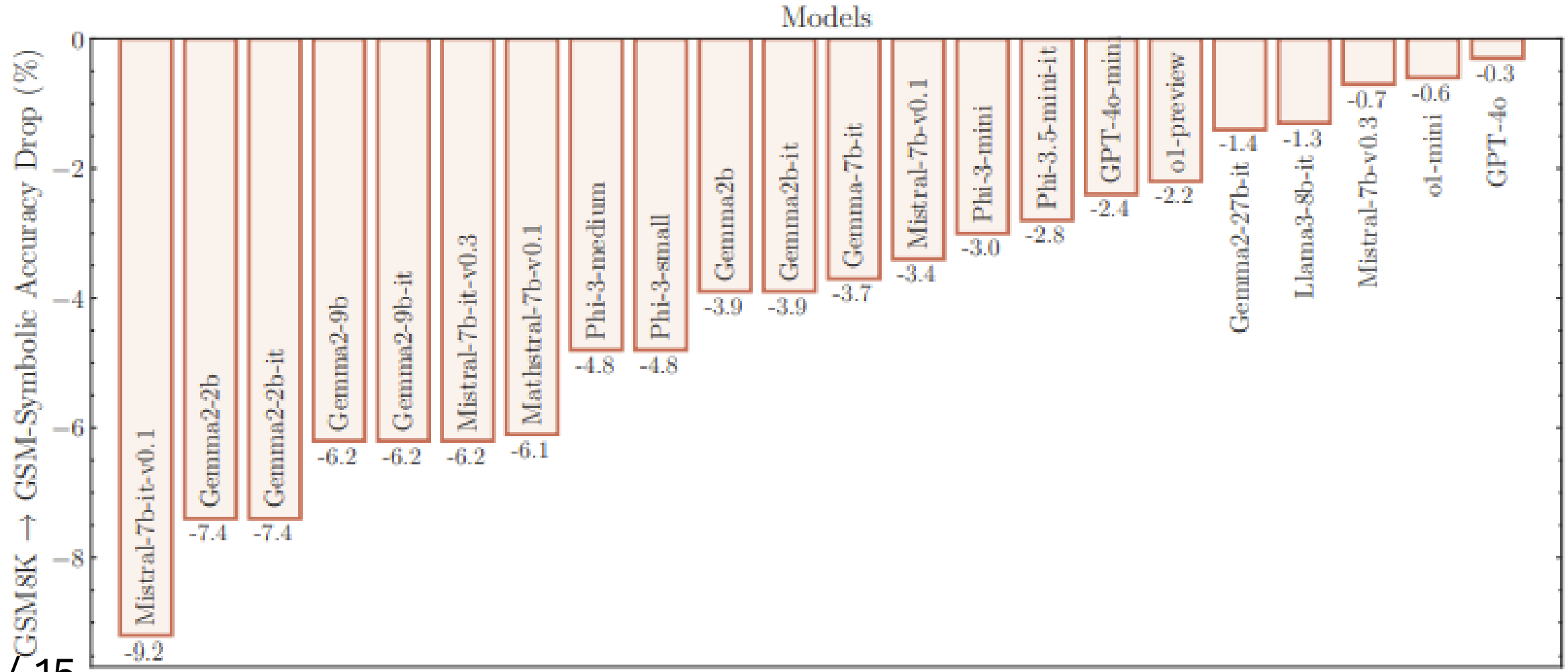
GSM Symbolic Template

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?

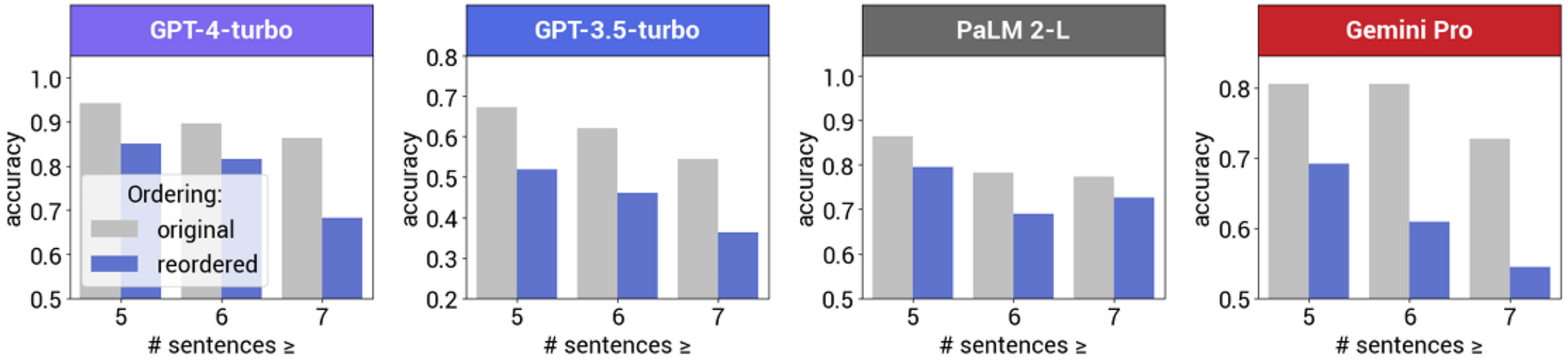
```
#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

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

https://arxiv.org/abs/2410.05229

有多少答案可能是「記憶」出來的？



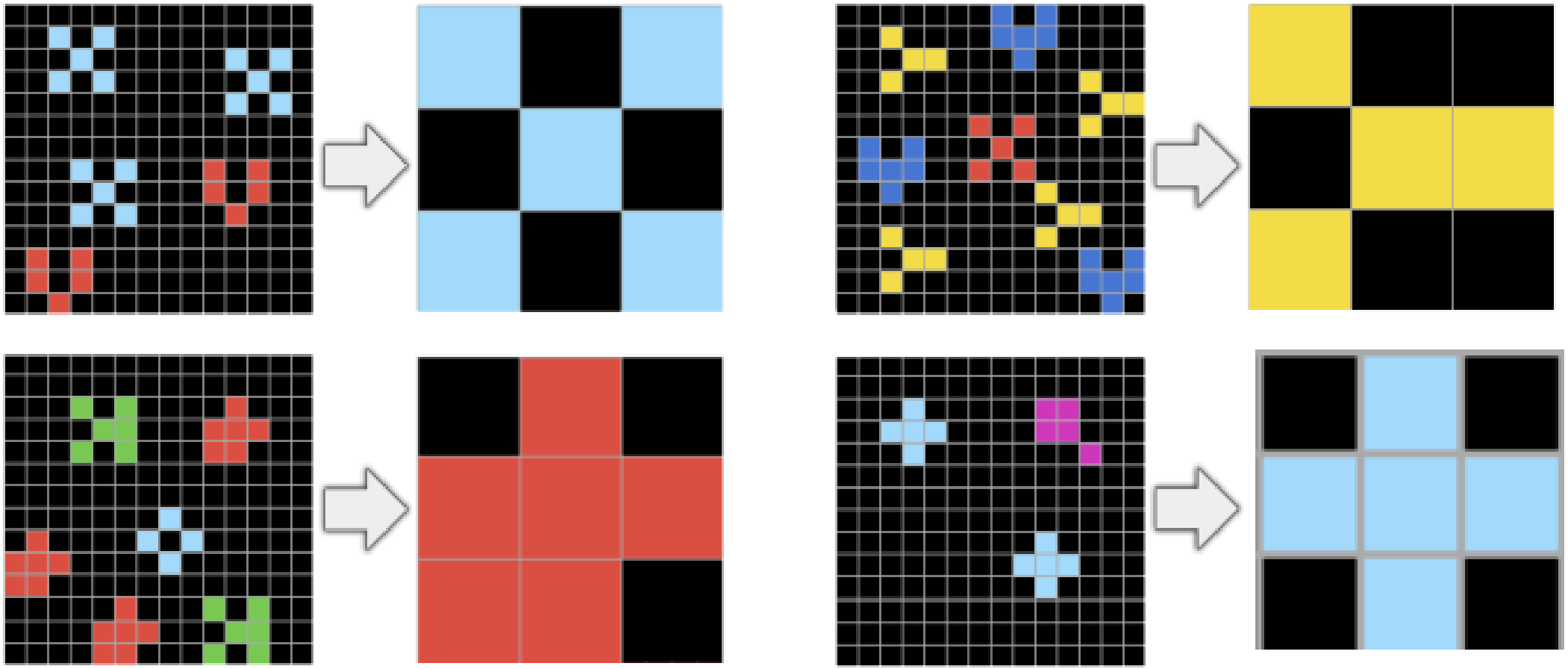
有多少答案可能是「記憶」出來的？



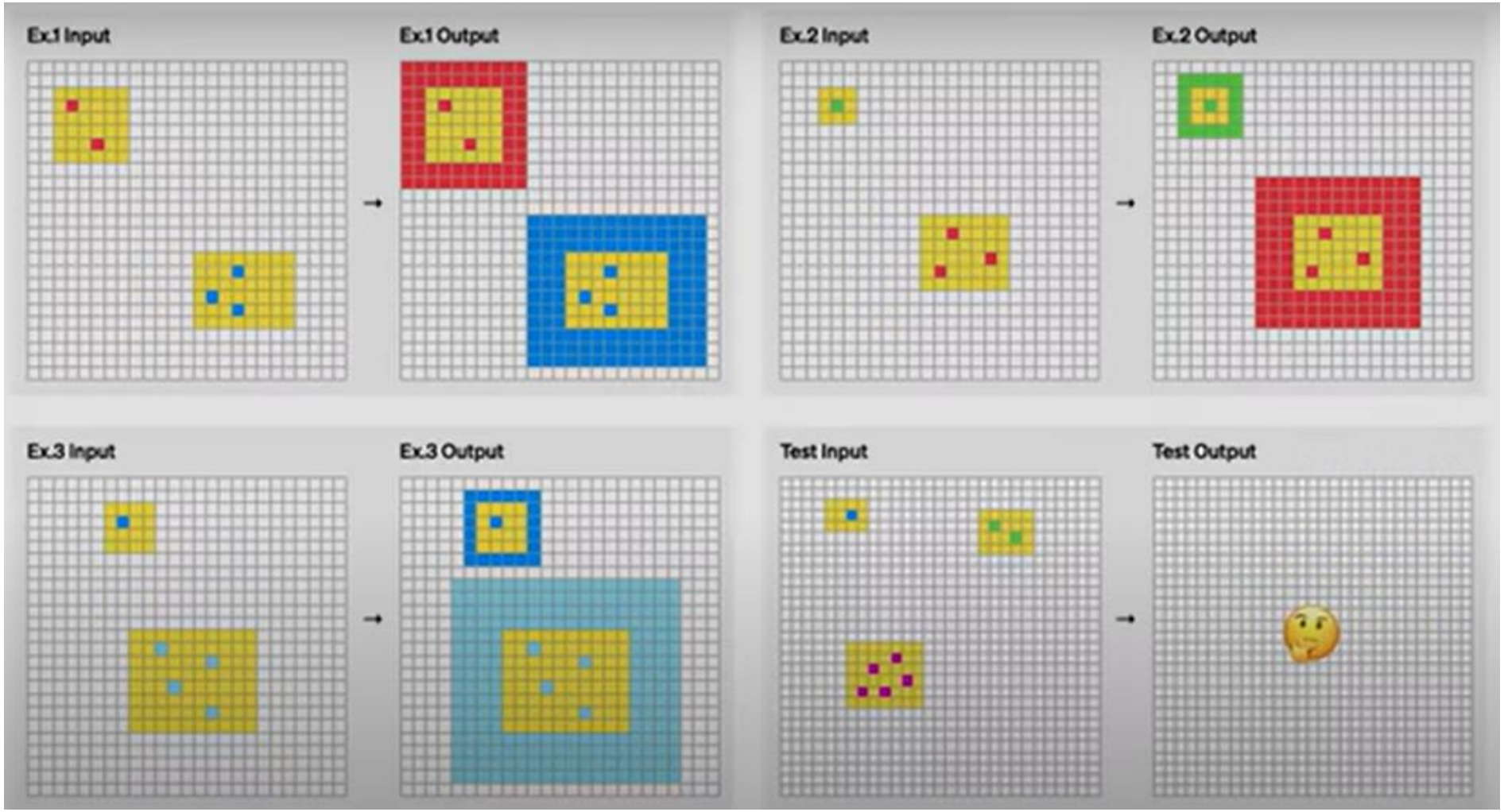
https://arxiv.org/abs/2402.08939

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
```

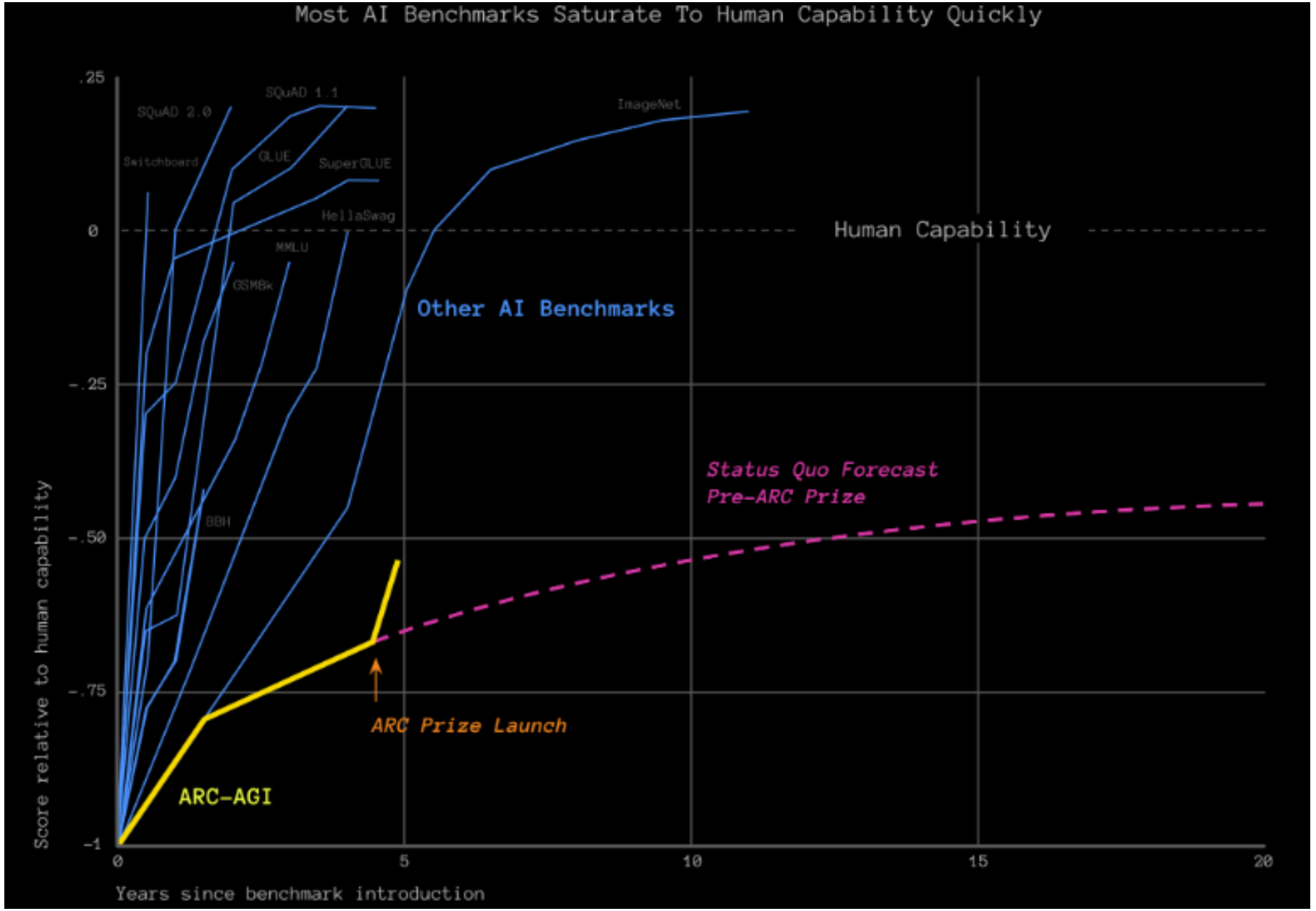
Output:

```
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
```

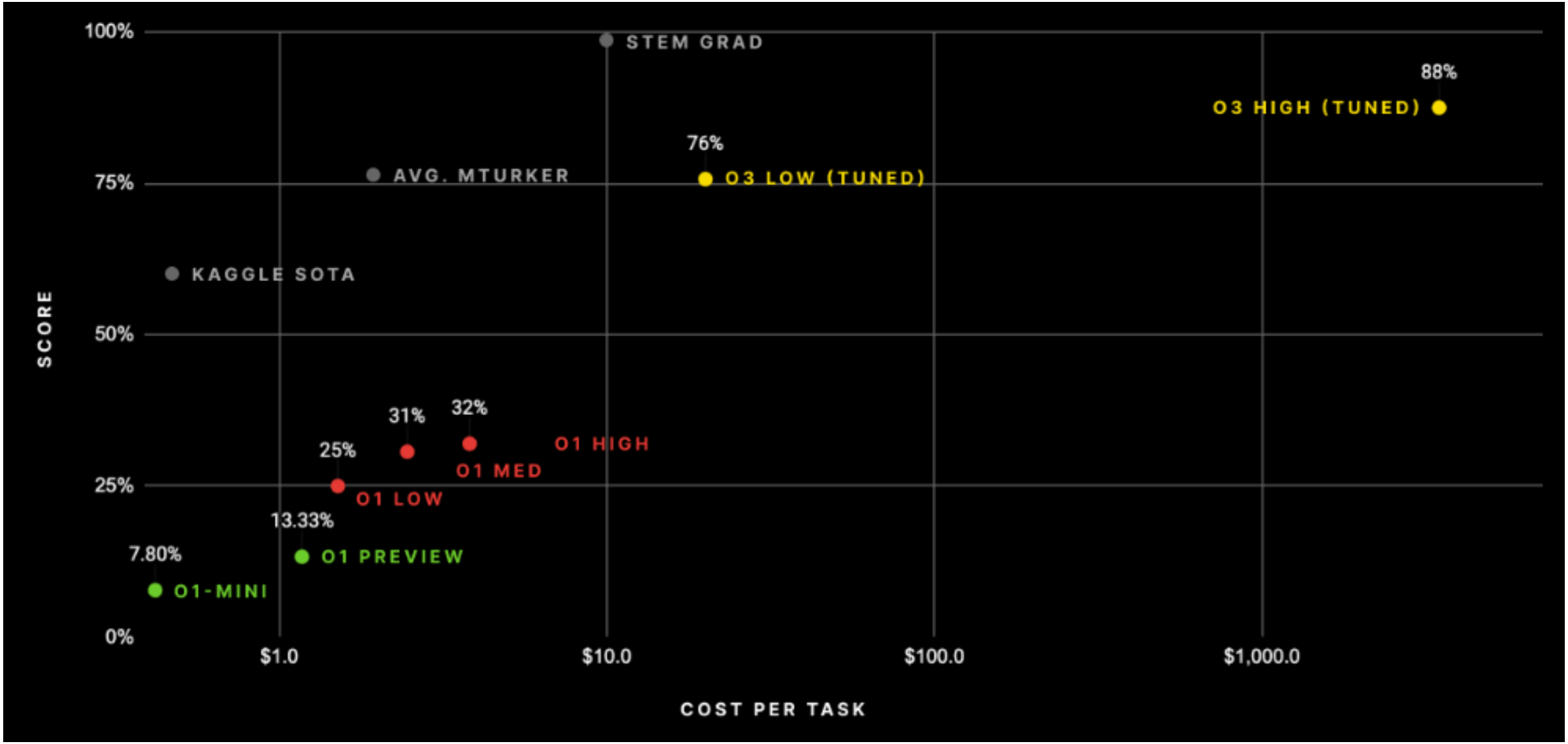
Input:

```
0 4 0
0 0 0
4 0 0
```

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



https://arcprize.org/blog/3-month-update



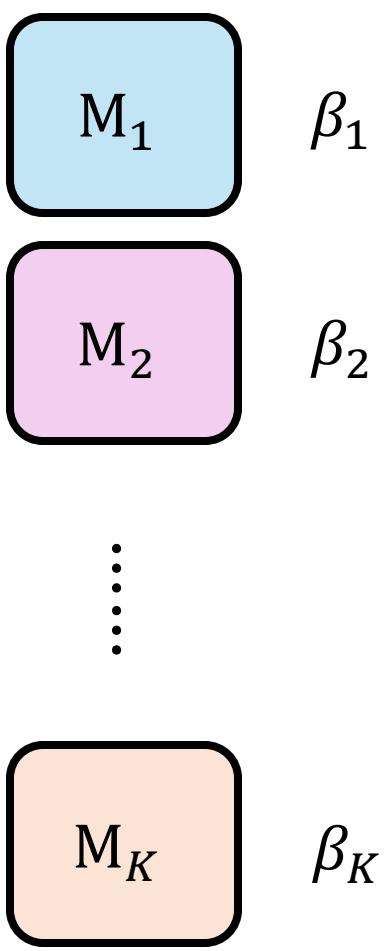
https://arcprize.org/blog/oai-o3-pub-breakthrough

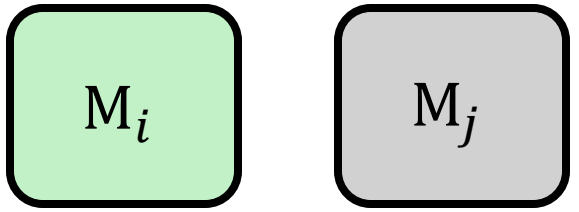
Chatbot Arena

https://lmarena.ai/

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

Chatbot Arena - Elo Score





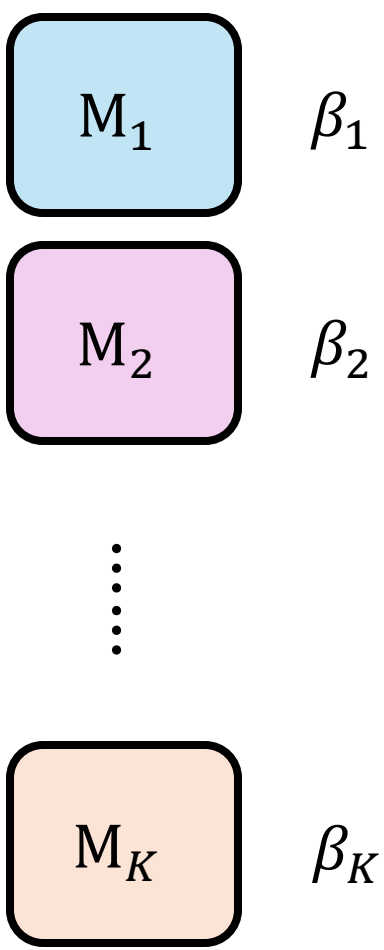
$$\frac{1}{1 + exp\left(-\frac{\beta_i - \beta_j}{400}\right)} = E_{i,j}$$

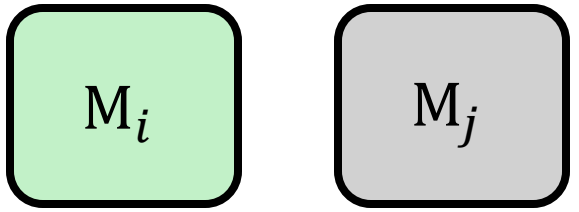
根據比賽結果統計勝率

算出 $\beta_1, \beta_2, \ldots, \beta_K$

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

Chatbot Arena - Elo Score





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

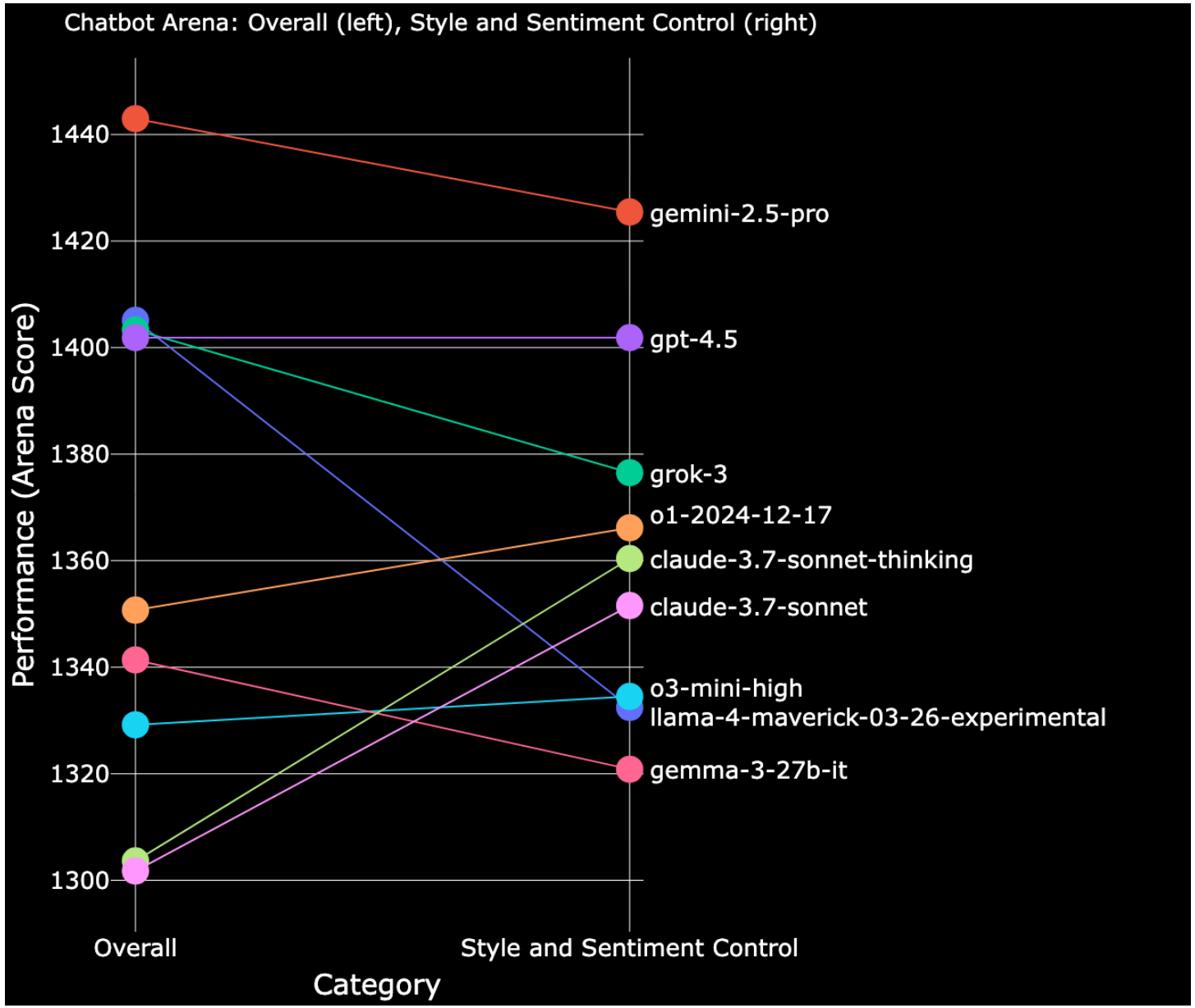
根據比賽結果統計勝率

算出 $\beta_1,\beta_2,\dots,\beta_K$

算出 $\gamma_1,\gamma_2,\dots$

β_0 模型實力以外的因素

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



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。