# FrugalGPT:
# 來看看窮人怎麼用 ChatGPT

Frugal: 節儉

# FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance

Lingjiao Chen, Matei Zaharia, James Zou

Stanford University

https://arxiv.org/abs/2305.05176

# GPT4 的 API 也是很花錢的

- 輸入：0.03$ / 1000 tokens
- 輸出：0.06$ / 1000 tokens

假設每次使用輸入 1000 tokens 、輸出 1000 tokens

每次使用需要 0.03$ + 0.06$ = 0.09$ (2.78 新台幣)

桃市府試驗以ChatGPT分析1999陳情案件

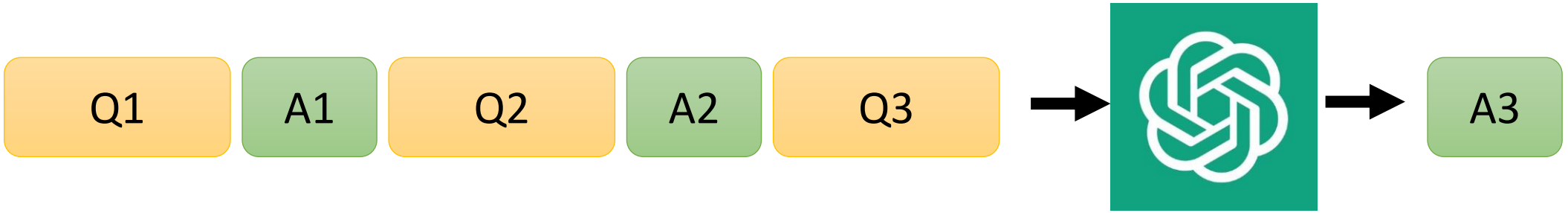
https://news.ltn.com.tw/news/politics/breakingnews/4273981

「1999臺北市民當家熱線」平均每月服務 15 萬 7,522 通電話

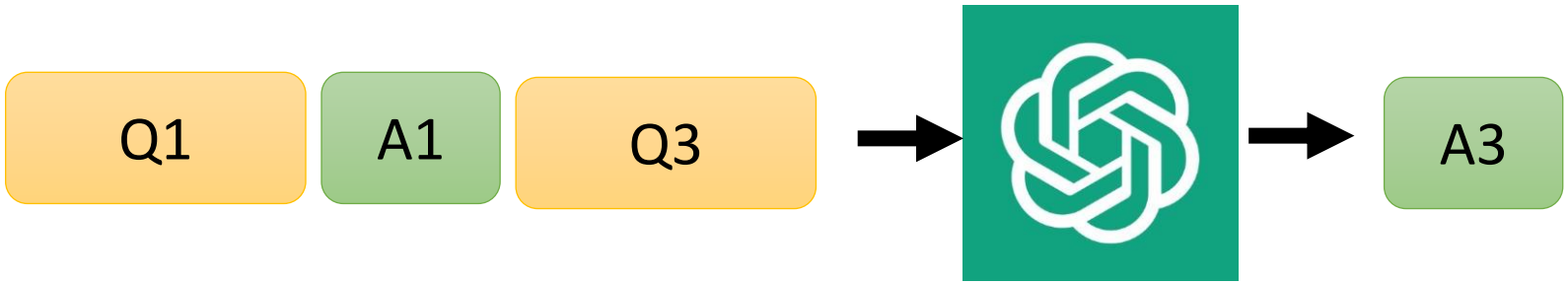
https://rdec.gov.taipei/cp.aspx?n=EE54BD6678096F88

# 方法一: Prompt Adaptation (縮短輸入)

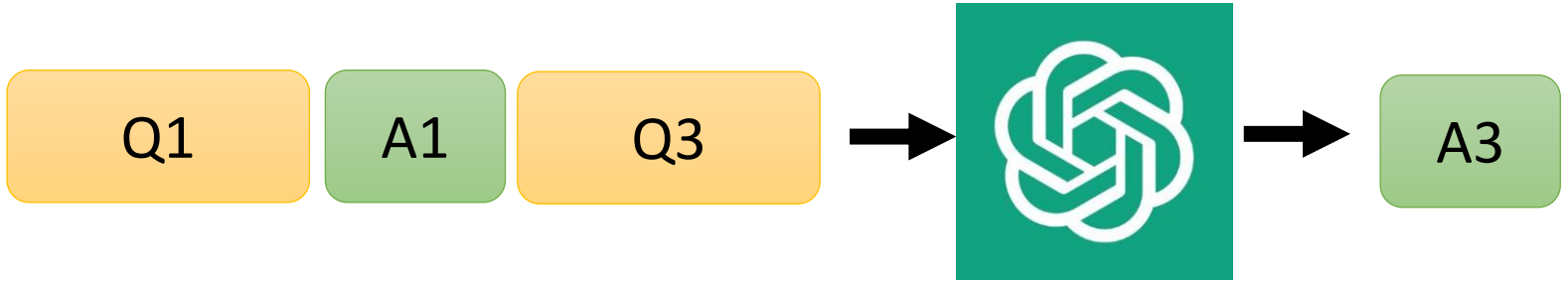
- In-context learning

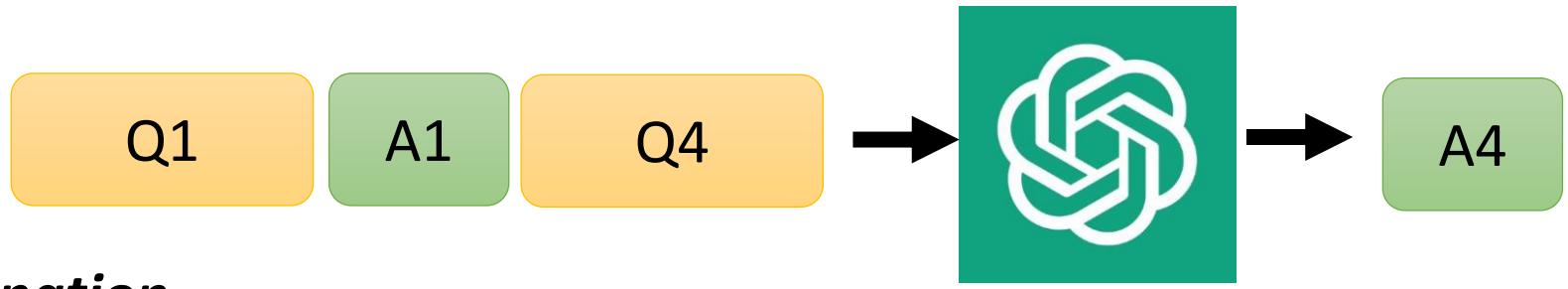



***Prompt Selection***

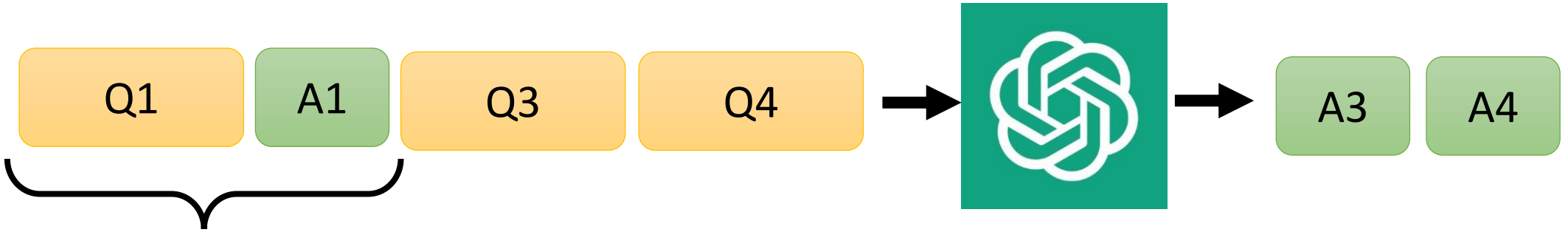


# 方法一: Prompt Adaptation (縮短輸入)




***Query Concatenation***

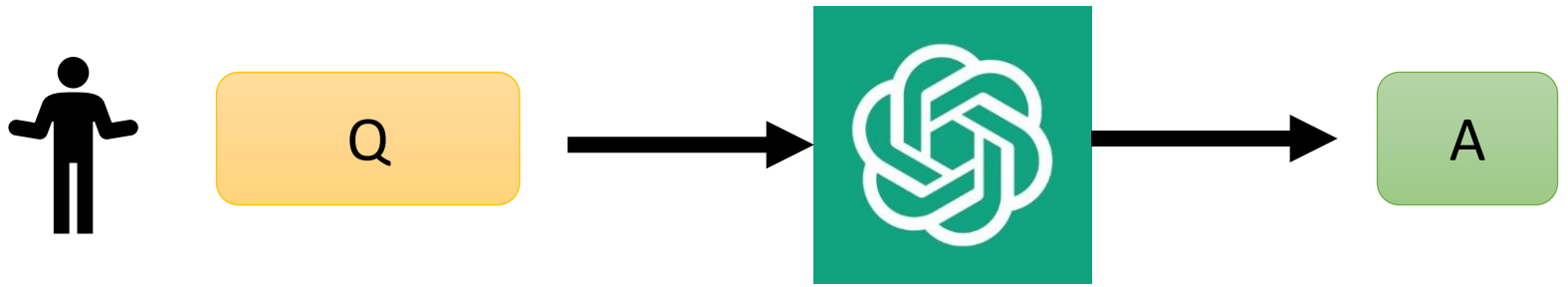


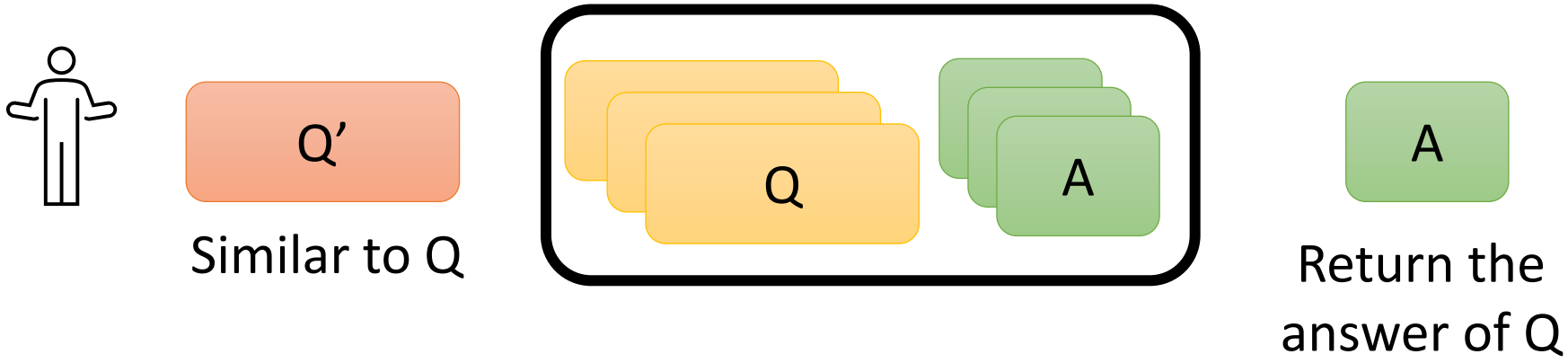
# 方法二: LLM Approximation (自建模型)


【生成式AI】窮人如何低資源復刻自己的 ChatGPT https://youtu.be/rK_rZFew1yc

# 方法二: LLM Approximation (自建模型)

- Completion Cache



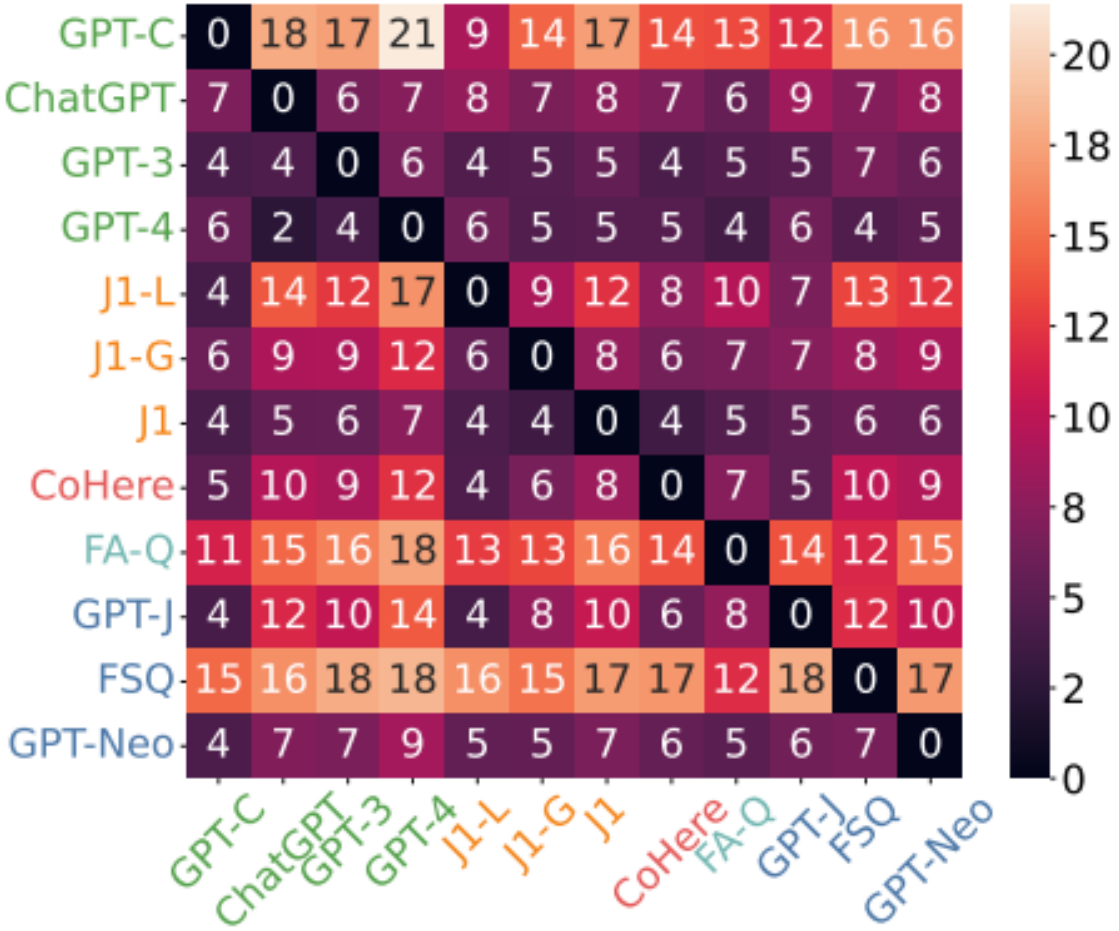



# 方法三: LLM cascade

| Provider | API | Size/B | Cost (USD) | | |
|---|---|---|---|---|---|
| | | | 1M input tokens | 1M output tokens | request |
| OpenAI | GPT-Curie | 6.7 | 2 | 2 | 0 |
| | ChatGPT | NA | 2 | 2 | 0 |
| | GPT-3 | 175 | 20 | 20 | 0 |
| | GPT-4 | NA | 30 | 60 | 0 |
| AI21 | J1-Large | 7.5 | 0 | 30 | 0.0003 |
| | J1-Grande | 17 | 0 | 80 | 0.0008 |
| | J1-Jumbo | 178 | 0 | 250 | 0.005 |
| Cohere | Xlarge | 52 | 10 | 10 | 0 |
| ForeFrontAI | QA | 16 | 5.8 | 5.8 | 0 |
| Textsynth | GPT-J | 6 | 0.2 | 5 | 0 |
| | FAIRSEQ | 13 | 0.6 | 15 | 0 |
| | GPT-Neox | 20 | 1.4 | 35 | 0 |

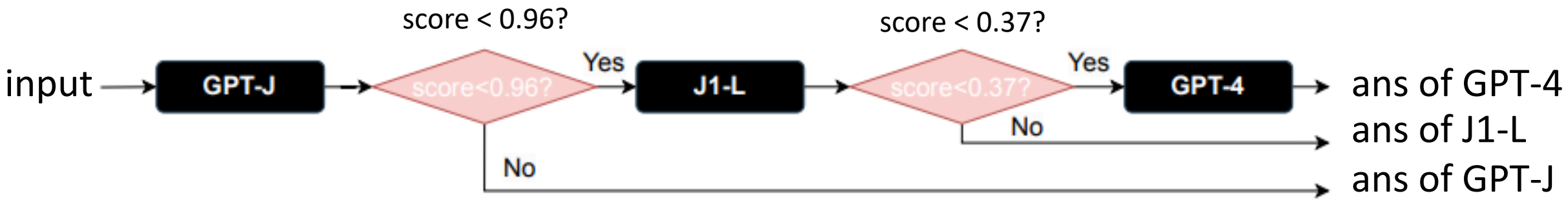
# 方法三: LLM cascade

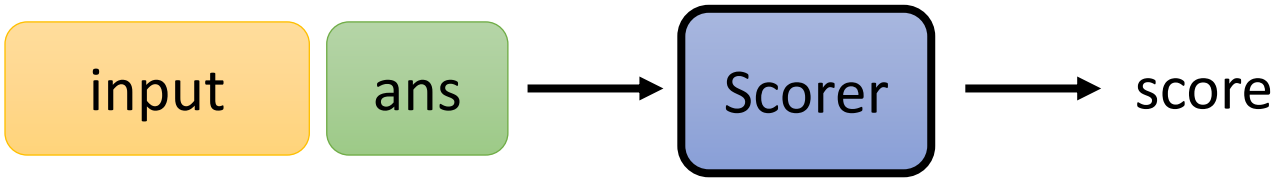
- 殺雞不用牛刀
  - 簡單的問題交給比較弱 (比較便宜) 的模型
  - 只有難的問題才給比較強 (比較貴) 的模型
- 不同模型的能力可能可以互補




HEADLINES

# 方法三: LLM cascade



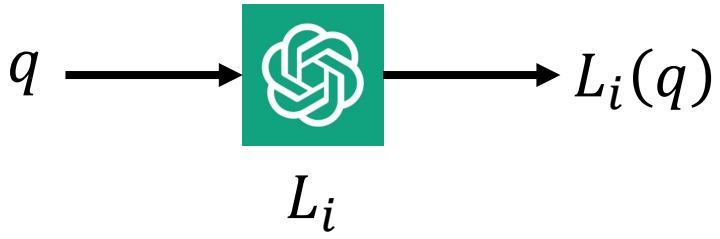

(需要另外訓練)

| Approch | Accuracy | Cost ($) |
|---|---|---|
| GPT-4 | 0.857 | 33.1 |
| FrugalGPT | **0.872** | **6.5** |

# 方法三: LLM cascade

$$\mathcal{L} = \{L_1, L_2, \dots, L_k\}$$

$$\mathcal{T} = \{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\}$$




$r\big(a, L_i(q)\big)$: performance of $L_i$

$s\big(q, L_i(q)\big)$: score from scorer

$$\max_{\mathcal{L},\mathcal{T}} \sum_{q,a} r\big(a, L_z(q)\big)$$

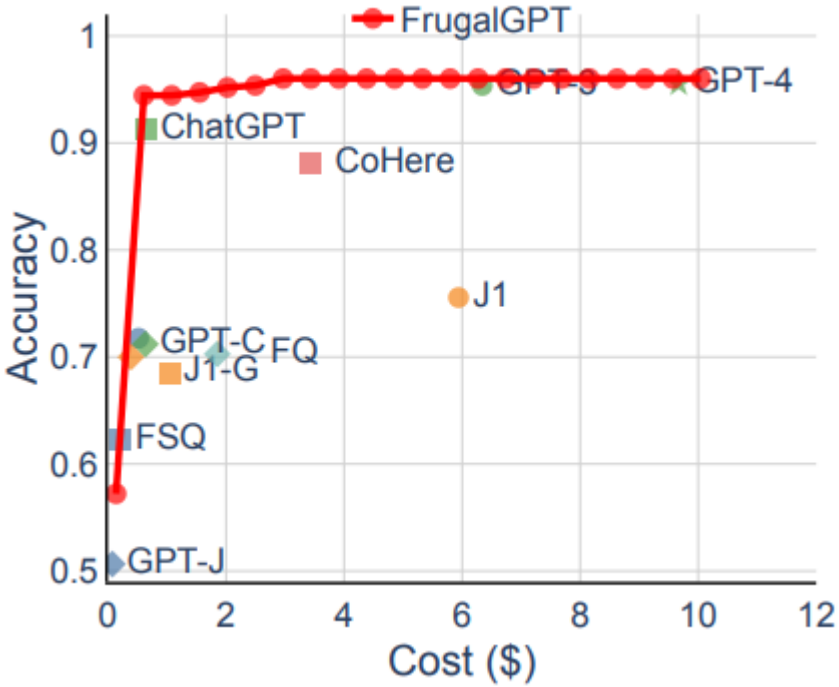
$z$ is the minimum $i$
such that $s\big(q, L_i(q)\big) > \lambda_i$

Cost of each request:

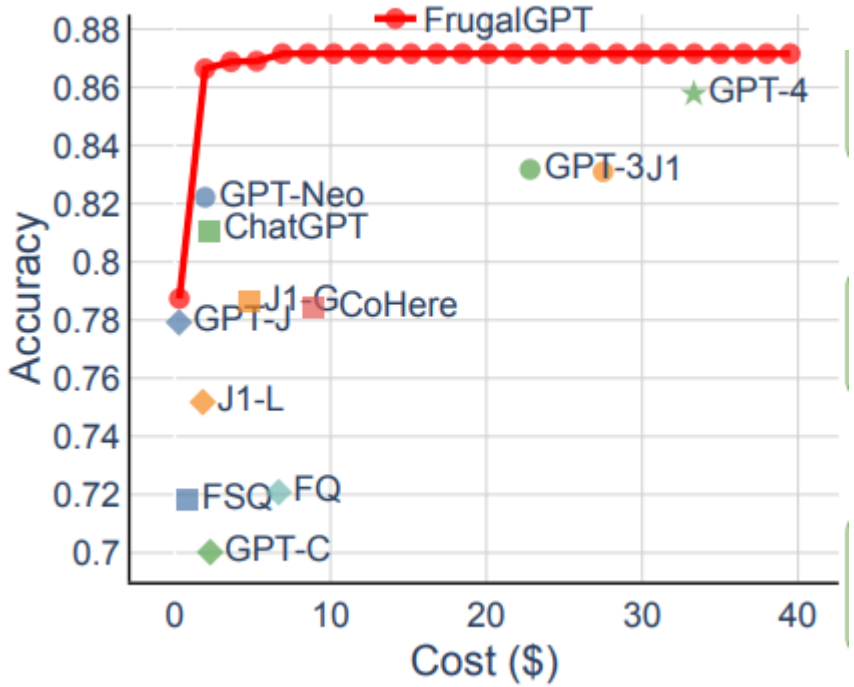
$$c_{L_i,1}\|q\| + c_{L_i,2}\|L_i(q)\| + c_{L_i,3}$$

$$\sum_{q}\sum_{i=1}^{z}\left[c_{L_i,1}\|q\| + c_{L_i,2}\|L_i(q)\| + c_{L_i,3}\right] < B$$

(budget)

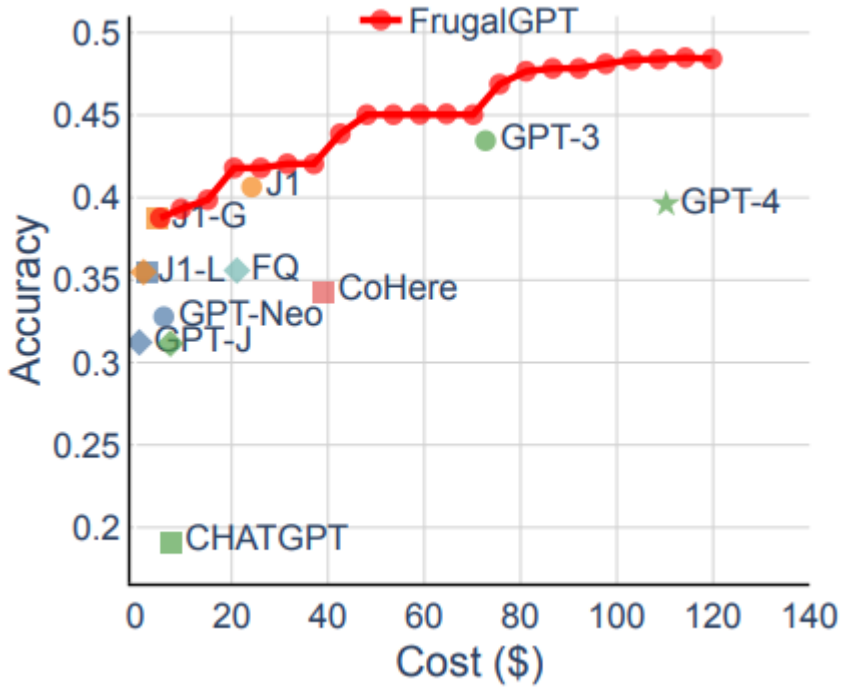



## HEADLINES




## OVERRULING

## COQA



# FrugalGPT:
# 來看看窮人怎麼用 ChatGPT

Frugal: 節儉