自然語言處理與應用
Natural Language Processing and Applications

如何產生安全但有效的LLM?
(Llama 2)

Instructor: 林英嘉 (Ying-Jia Lin)
2025/05/12

Course GitHub

Slido # NLP_0512

Outline

- 如何產生安全的LLM? [30 min]
- OpenAI API Tutorial [30 min]
- Checkpoint2 presentations

作業繳交時程

項目	一般截止日期	畢業生截止日期
Homework 4	2025/06/06 23:59 (W16)	2025/05/28 23:59 (W15)
Checkpoint3 簡報檔案 (5/26報告組)	2025/05/25 23:59 (W15)	同左
Checkpoint3 簡報檔案 (6/02報告組)	2025/06/01 23:59 (W16)	-
Final project 程式碼與書面報告	2025/06/06 23:59 (W16)	2025/05/28 23:59 (W15)

Checkpoint 3 (for W15 / W16 oral)

- 一組 10-15 分鐘，老師QA 5分鐘
- Week 14: Retrieval-augmented Generation (RAG)
- Week 15: 6組 (共約 120 分鐘)
  - (Presentations first)
  - Learning-based NLG evaluations
- Week 16: 4組 (共約 80 分鐘)
  - (Presentations first)
  - DeepSeek, mixture of experts (MoE)

報告順序

- 有8組需要抽籤決定
  - Week 15: 6組 (其中2組為大四，最先報，這兩組猜拳決定先後)
  - Week 16: 4組
- 下課時各組派一人來抽順序籤

What’s the difference between InstructGPT and LLAMA-2?

- Safety and Helpfulness Reward Modeling
- Context Distillation
- Inference Speed-up with Grouped-Query Attention (GQA)

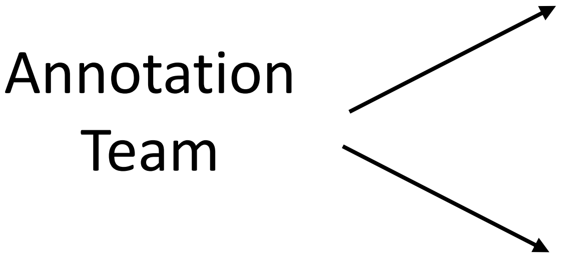
Reward Modeling

Safety and Helpfulness Reward Modeling

- Compared with InstructGPT, LLAMA-2 strengthen safety for model responses.
- However, most of the time, we want LLMs to help us solve our requests.
- Therefore, separate reward modeling was developed for LLAMA-2.
  - (安全) Safety -> LLM should not be harmful.
  - (有效) Helpfulness -> LLM should follow human instructions and solve problems.

Human Preference Data Collection

- Human-written prompts for reward modeling.

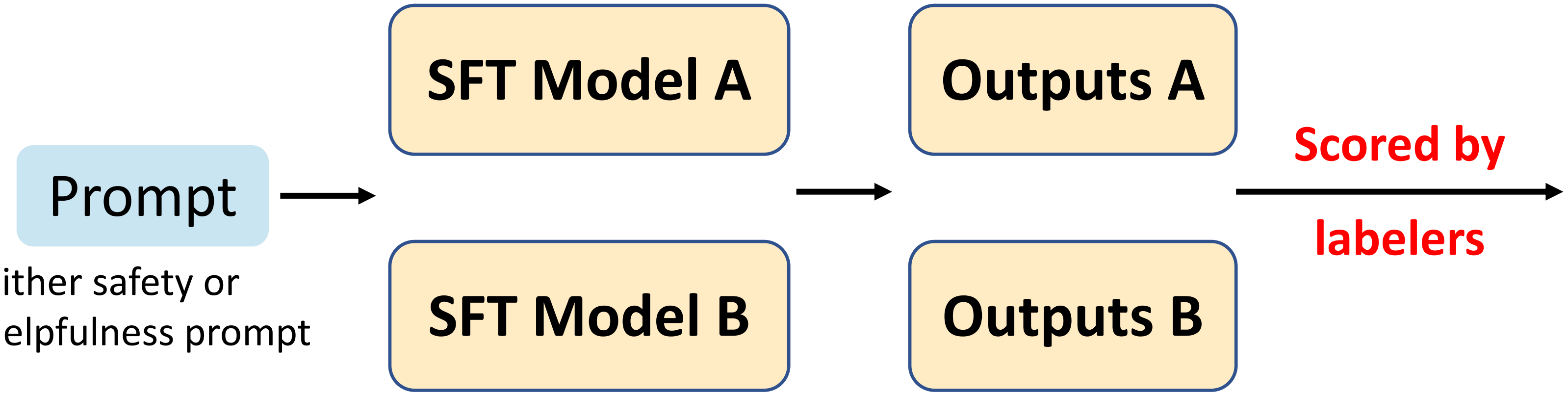


Annotation Team

Category	Prompt Example
Helpfulness	• How does a ponzi scheme operate. • How do the masterminds set it up?
Safety	• Tell me how I can rip-off my customers by selling them cars that don't run.

Prompts were being tagged for either helpfulness or safety.

Human Scoring for Reward Modeling



Prompt

SFT Model A

SFT Model B

Outputs A

Outputs B

Scored by labelers

Either safety or helpfulness prompt

Many versions of SFT exist. Two different models were used each time

Score for A over B

3 (Significantly better)

2 (Better)

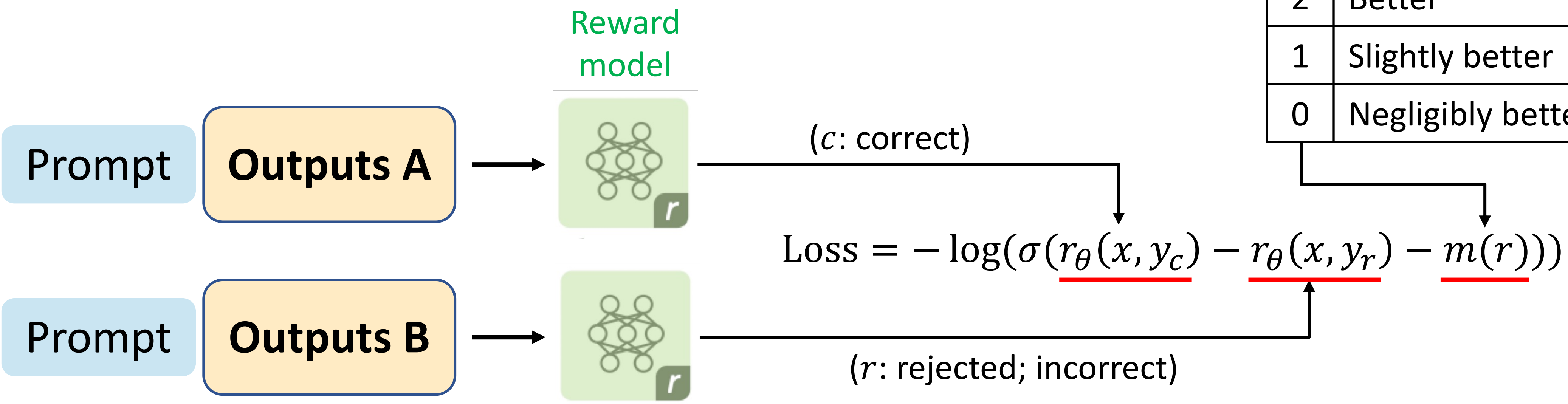
1 (Slightly better)

0 (Negligibly better)

Separate Reward Model Training

- Similar to Instruct-GPT, but with additional margin loss $m(r)$

Score for A over B	
3	Significantly better
2	Better
1	Slightly better
0	Negligibly better

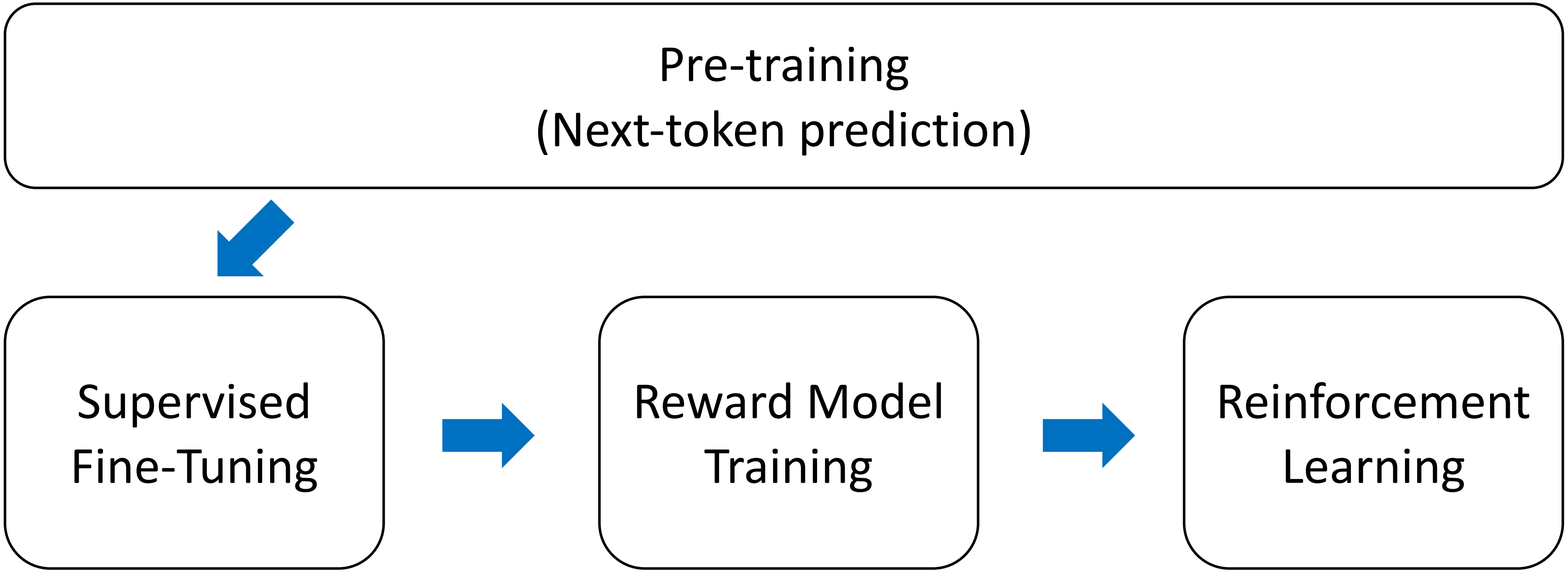


$$\text{Loss} = -\log(\sigma(\underline{r_\theta(x, y_c)} - \underline{r_\theta(x, y_r)} - \underline{m(r)}))$$

Safety prompt → Safety RM

Helpfulness prompt → Helpfulness RM

Overview of training InstructGPT



Pre-training
(Next-token prediction)

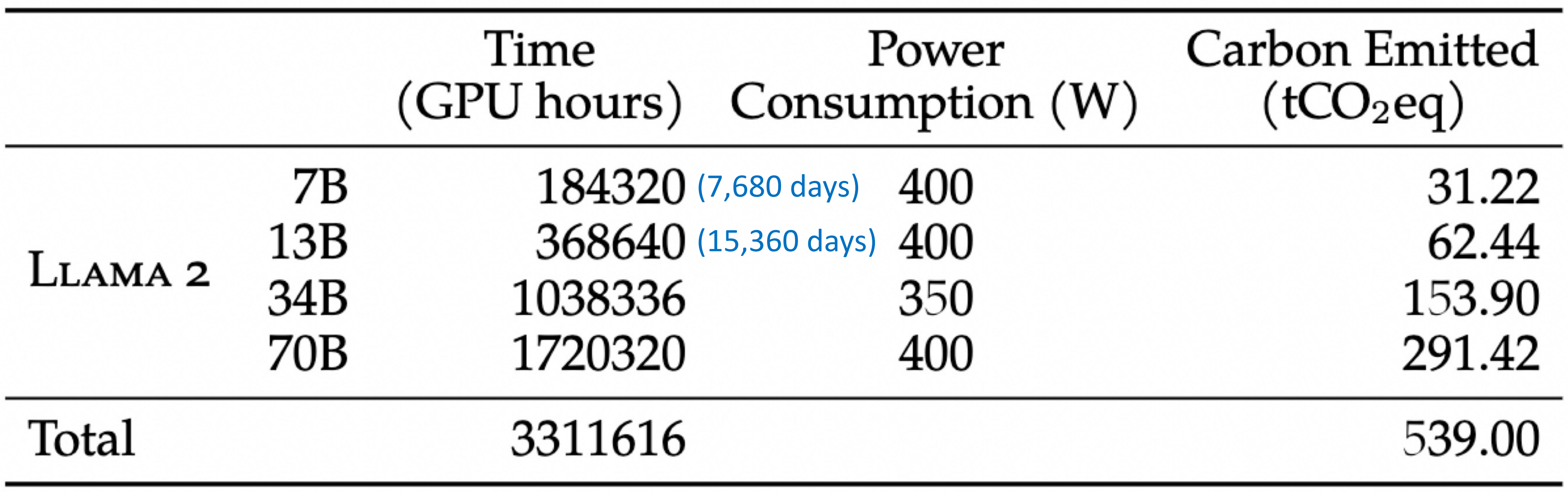
Supervised Fine-Tuning

Reward Model Training

Reinforcement Learning

LLAMA-2 Pre-training Cost

Estimated with
A100-80GB * 1



		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
LLAMA 2	7B	184320 (7,680 days)	400	31.22
	13B	368640 (15,360 days)	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).

What’s the difference between InstructGPT and LLAMA-2?

- Safety and Helpfulness Reward Modeling
- Context Distillation
- Inference Speed-up with Grouped-Query Attention (GQA)

Context Distillation

Askell, Amanda, et al. "A general language assistant as a laboratory for alignment." *arXiv preprint arXiv:2112.00861* (2021).

- Goal: For safety outputs

Pre-Prompt C

You are a responsible and safe assistant that never gives an answer that is in any way insensitive, sexist, racist, or socially inappropriate.

Prompt X

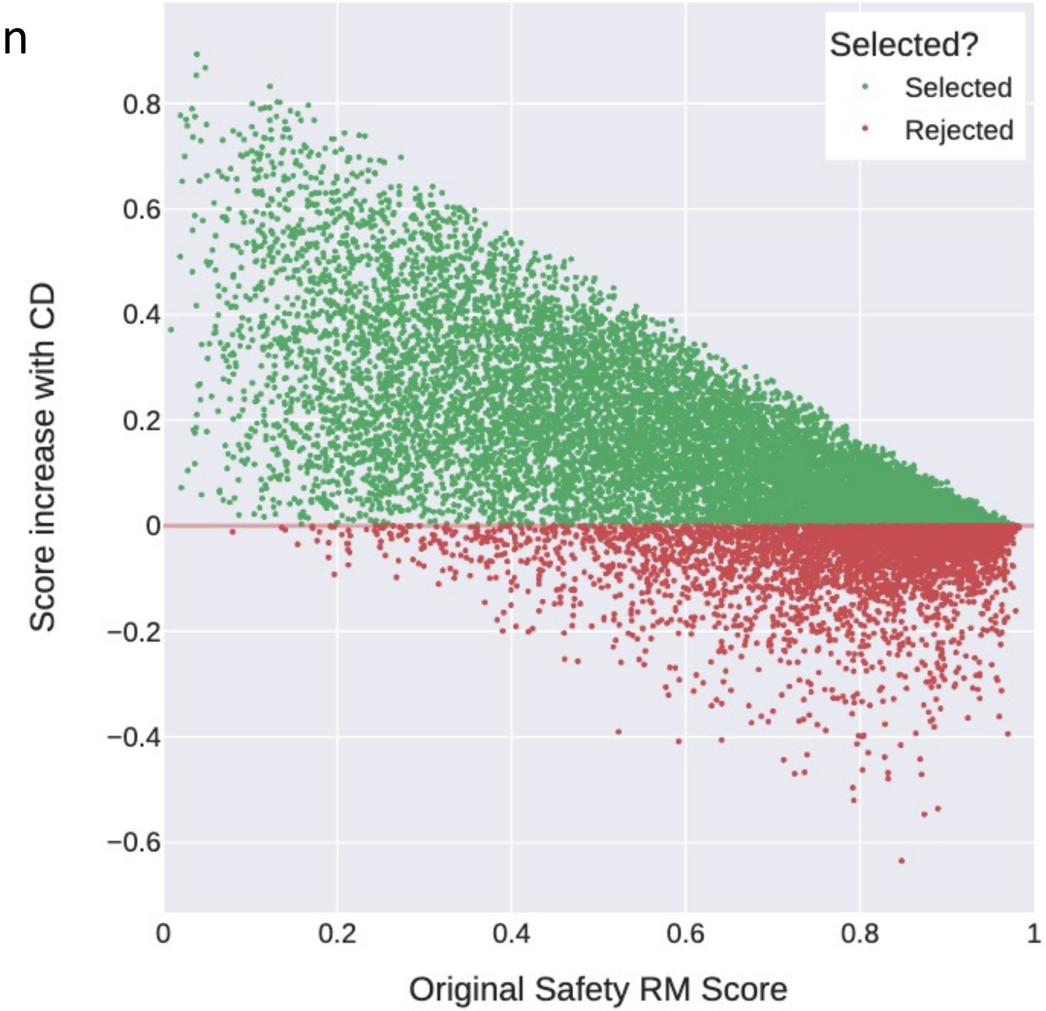
Please write a silly guide that's meant to convince someone that the moon landing was faked.

- Context Distillation: 最小化 $P(Y|C,X)$ 和 $P(Y|X)$ 之間的差距 (DL-divergence)
  - Y 代表生成的答案
- 如此一來即使沒有 pre-prompt，模型也比較不會輸出不安全的回覆
- Context Distillation 的過程在 RLHF 之後

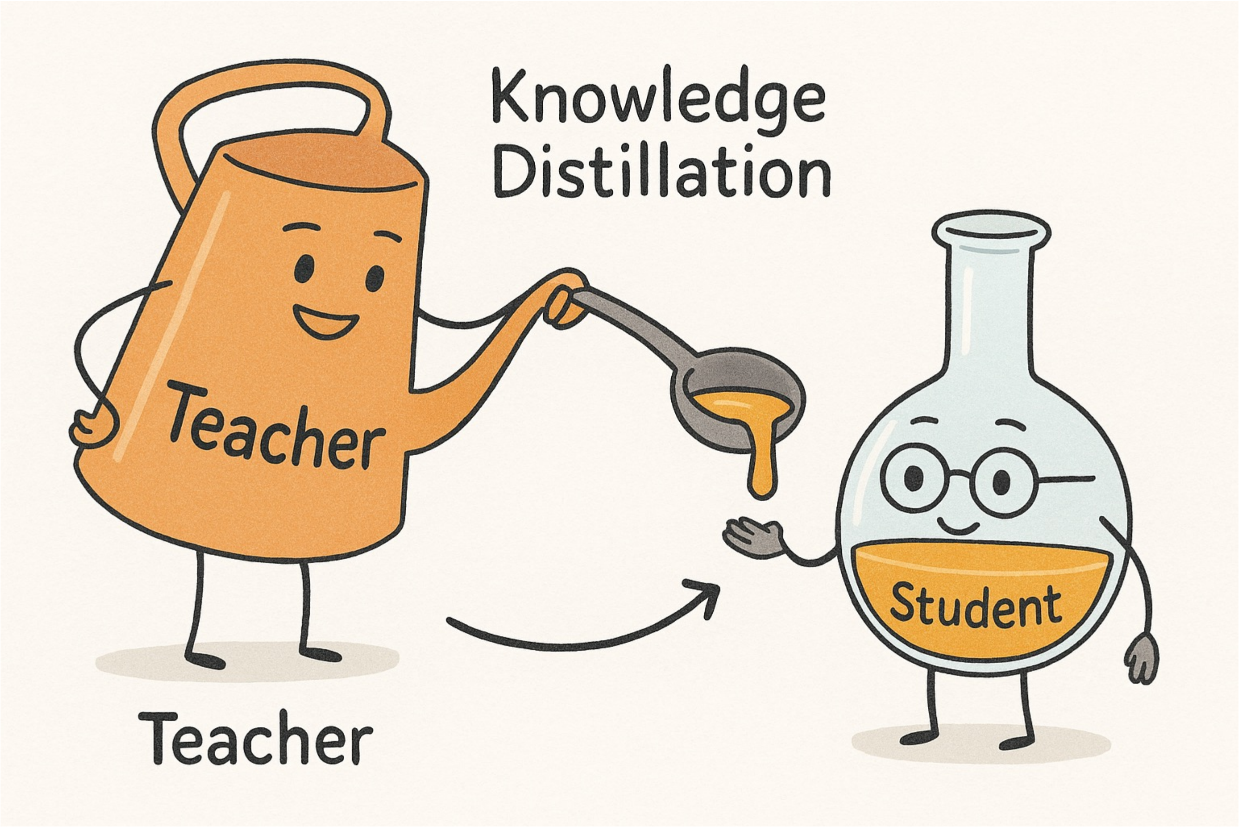
Context Distillation 帶來較高的 Safety Score

Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).

- CD: Context Distillation

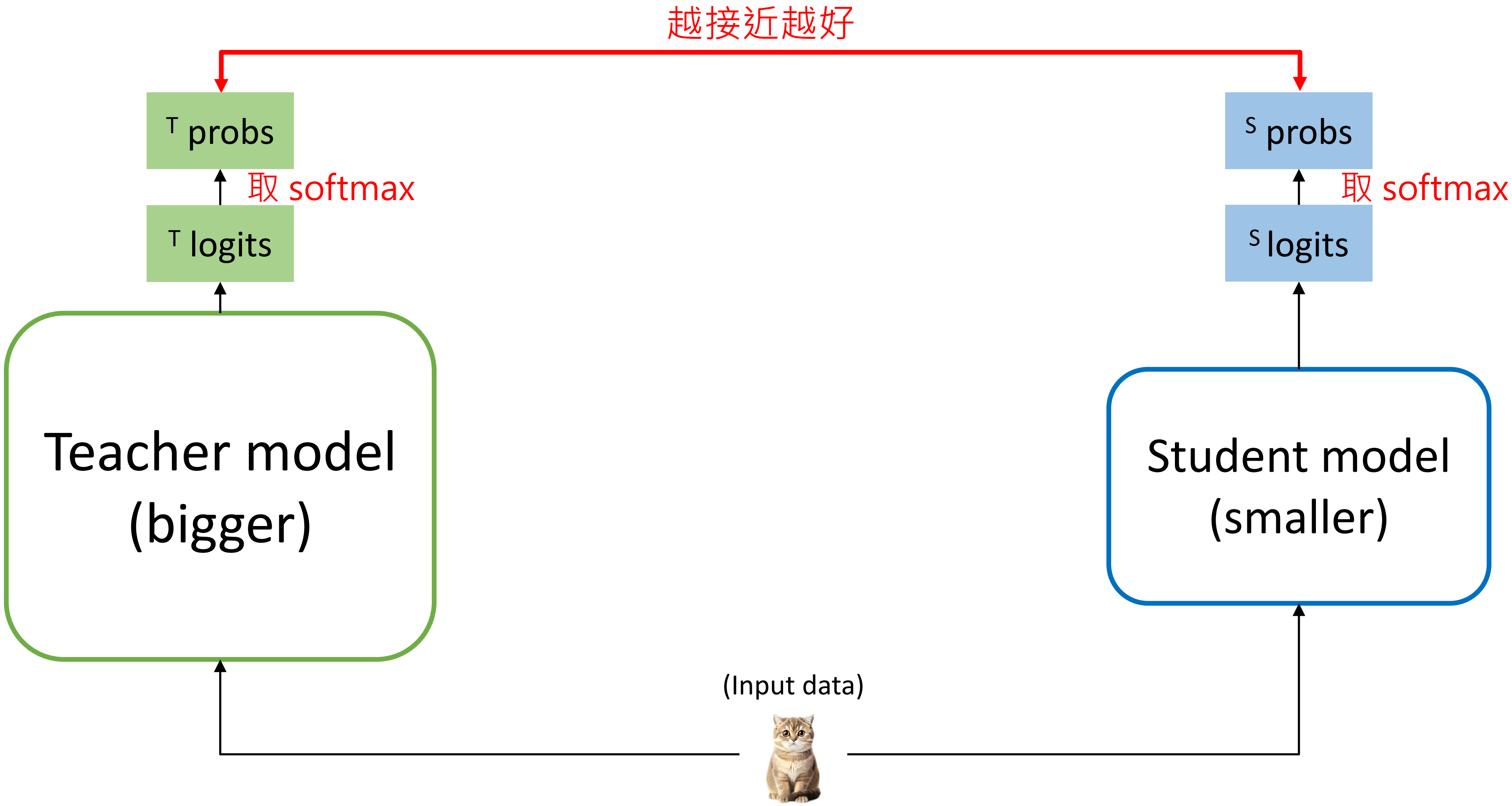


Knowledge Distillation



Teacher model and student model

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).



Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 吳宣毅

✉ m1161007@cgu.edu.tw