# 大型推理語言模型的訓練三步驟

**羅紹元**

助理教授、玉山青年學者

國立臺灣大學資訊工程學系

03/14/2026 @ 臺大醫院外科部

# Three Training Stages of Large Reasoning Language Models

Shao-Yuan Lo

Assistant Professor @ National Taiwan University

03/14/2026 @ NTUH

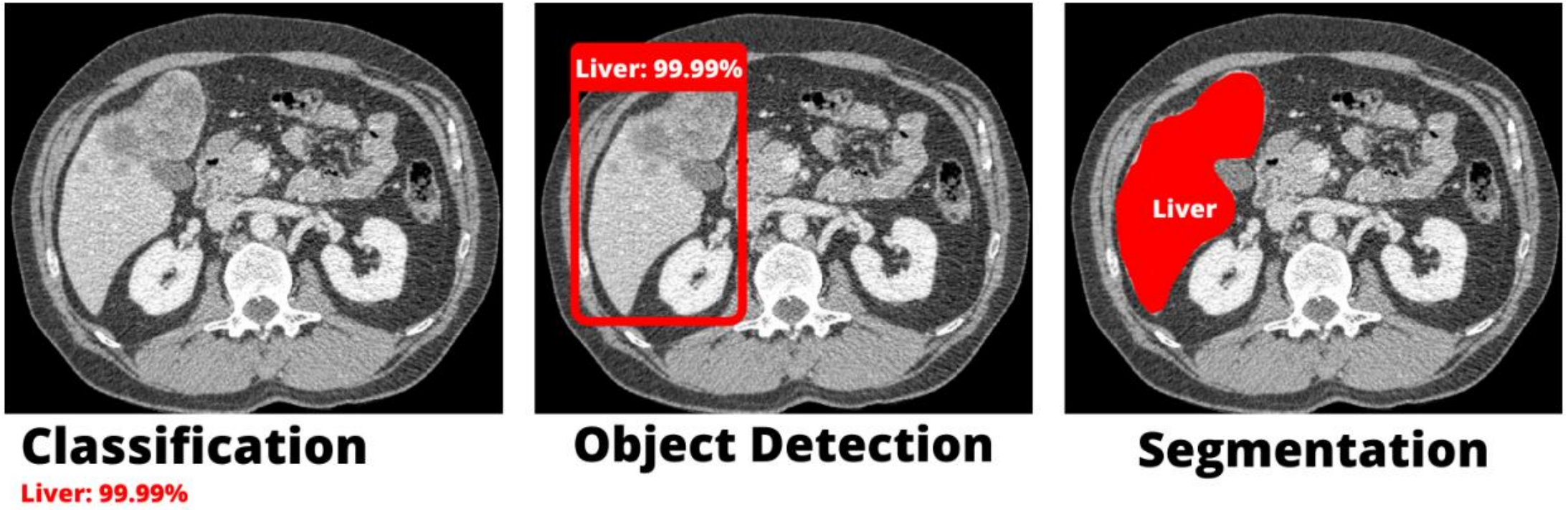
# 羅紹元 Shao-Yuan Lo

- Assistant Professor @ **National Taiwan University**
  Taiwan (2025 – Present)
- Research Scientist @ **Honda Research Institute USA**
  San Jose, CA (2023 – 2025)
- Research Intern @ **Amazon**
  Seattle, WA (Summer 2021 & 2022)

- PhD in ECE @ **Johns Hopkins University**
  Baltimore, MD (2019 – 2023)
- MS in EE @ **National Chiao Tung University**
  Taiwan (2017 – 2019)
- BS in EECS @ **National Chiao Tung University**
  Taiwan (2013 – 2017)

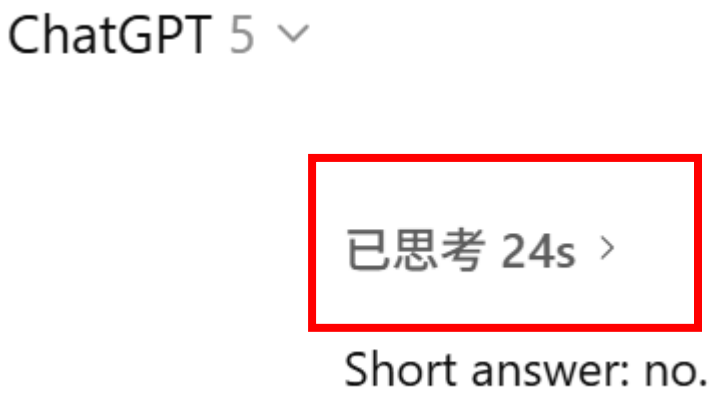
# What Can Be Done Without LLMs?

- Input → Output (input-output mapping)
- Task-specific perception, fixed output format, not explainable



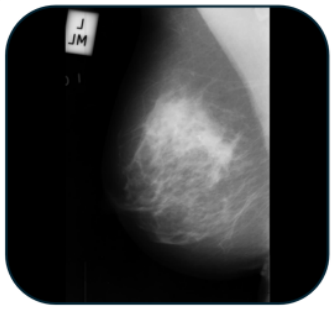

[Source credit: https://github.com/aminbenmansour/automatic-liver-segmentation]

# What Are Reasoning LLMs?

# What Reasoning LLMs Enable?

- Input → Reasoning → Output (thinking to answer)
- Flexible tasks, open-ended, natural language interaction, explainable



**Problem:** What does this image depict in terms of its content?

A)Brain tissue
B)Intestinal tissue
C)Kidney tissue
D)Breast tissue

**Med-R1 Response:**

**Think:** The image appears to be a mammogram, which is a type of X-ray used to detect abnormalities in the breast tissue. The presence of dense, irregular patterns and a central area of lighter density suggest the presence of breast tissue.

**Answer: D**

Groundtruth: D

**Problem:** What is the specific abnormality present in this image?

A)Pneumonia
B)Sinusitis
C)Urolithiasis
D)Osteoporosis

**Med-R1 Response :**

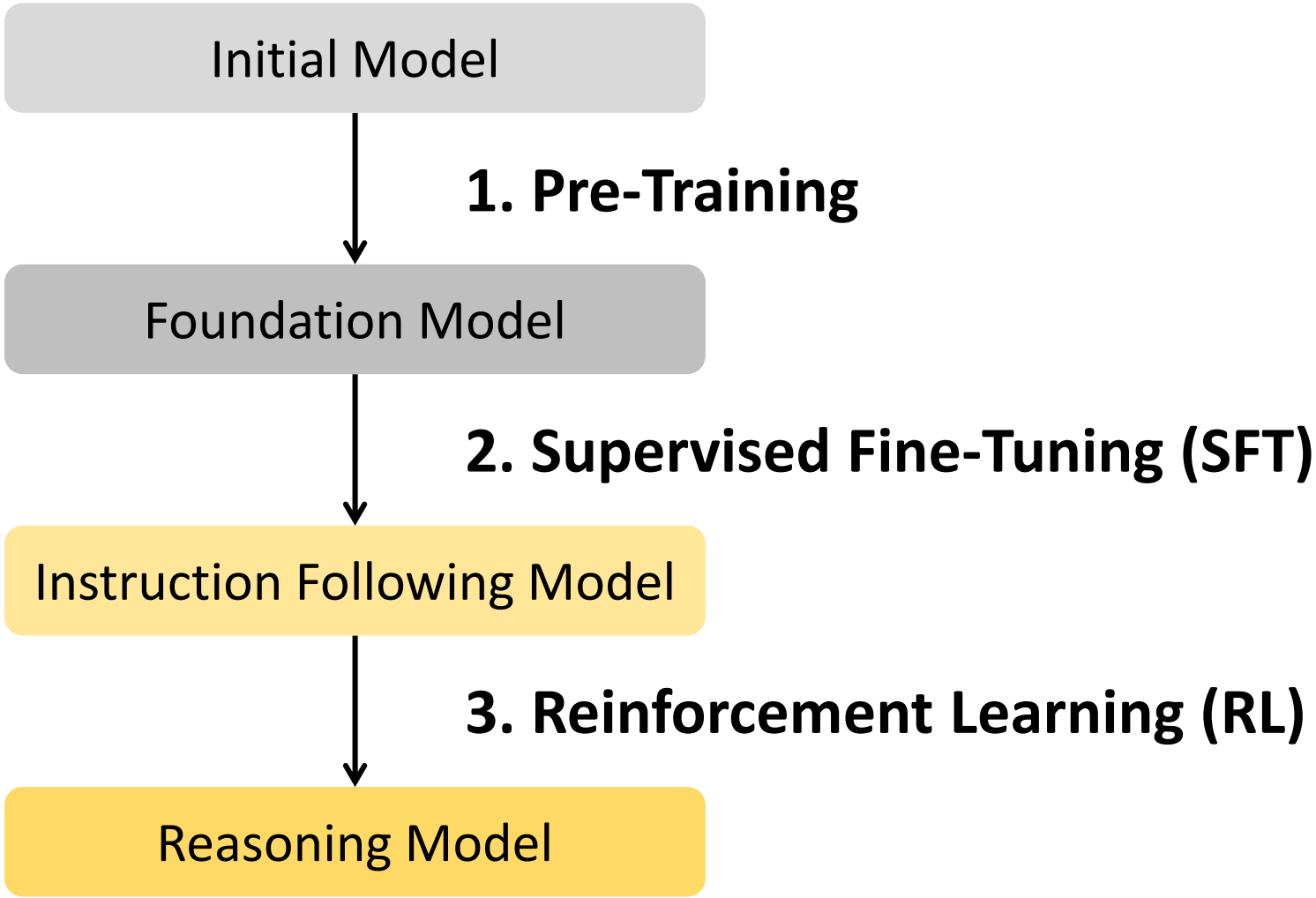
**Think:** The image is a computed tomography (CT) scan of the abdomen, which is a medical imaging technique used to visualize internal structures. The specific abnormality present in this image is not immediately clear from the visible structures alone. However, the presence of a large, dense mass in the right lower quadrant of the abdomen could suggest a urolithiasis.

**Answer: C**

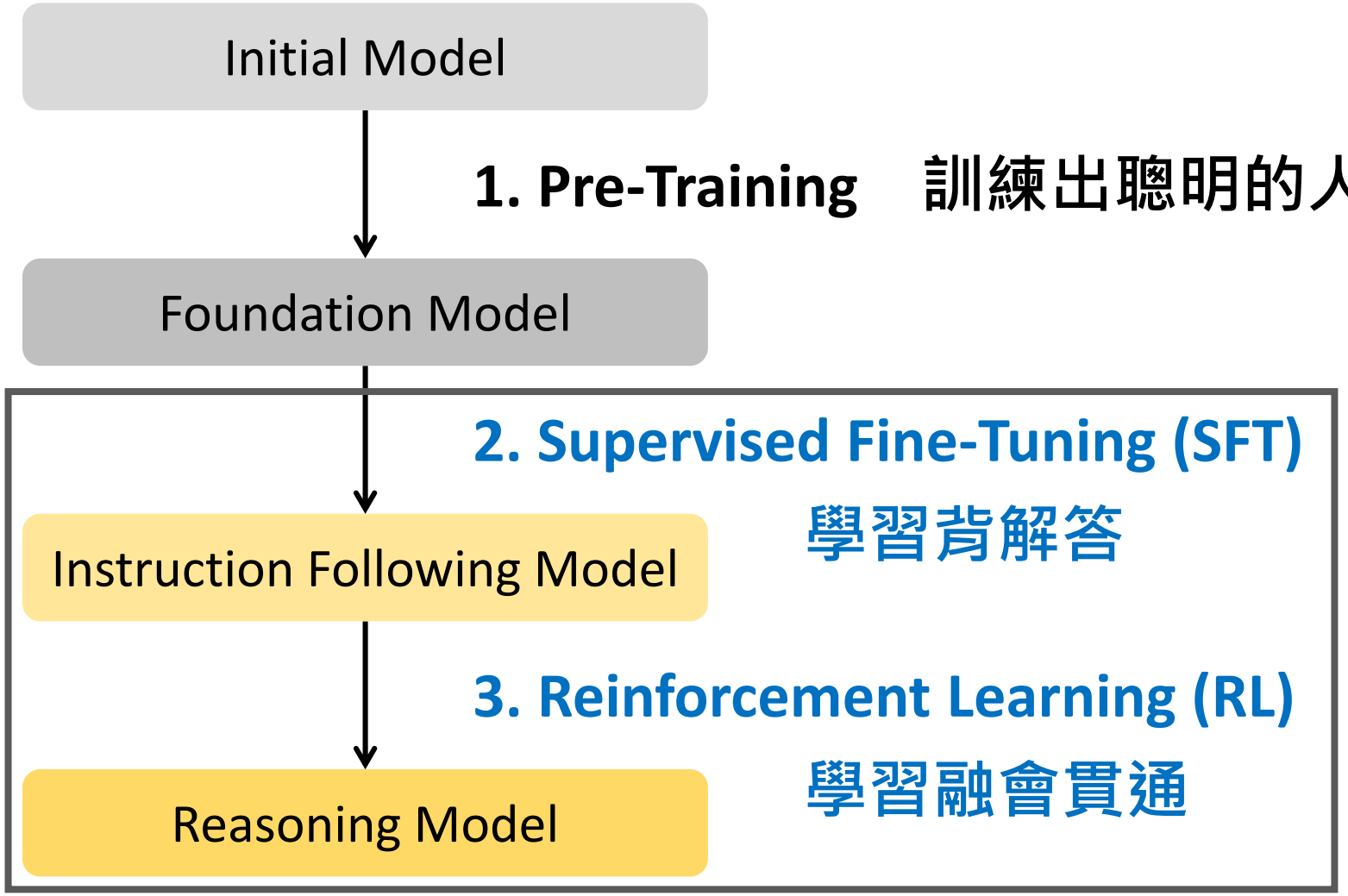
Groundtruth: C

[Lai et al. 2025]

# So, How to Train a Reasoning LLM? 3 Stages!



# So, How to Train a Reasoning LLM? 3 Stages!




**Post-Training**

將人才訓練成專家

# 1. Pre-Training

- Next-token prediction 文字接龍

$$P_\theta(x_t \mid x_{<t})$$

where $x_{<t} = (x_1, \ldots, x_{t-1})$

Training objective:

$$P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{<t})$$

我參加外科資料科學研討會

我參__
我參加__
我參加外__
我參加外科__
我參加外科資__
我參加外科資料__
我參加外科資料科__
我參加外科資料科學__
我參加外科資料科學研__
我參加外科資料科學研討__

# 1. Pre-Training

- Formulates all types of knowledge as a next-token prediction problem (autoregressive modeling)

| **Task** | **Example** |
|---|---|
| Grammar | In my free time, I like to __. {**run**, banana} |
| Lexical semantics | I spread butter on warm __. {**bread**, pencil} |
| World knowledge | The tallest mountain in the world is __. {**Mount Everest**, Mount Fuji}. |
| Sentiment analysis | Restaurant review: The food was delicious and the service was friendly. The experience was __. {**positive**, negative} |
| Translation | The word for “cat” in French is __. {**chat**, perro} |
| Spatial reasoning | The ball rolled under the __. {**table**, sky} |
| Logical reasoning | If it rains, the ground gets wet. If the ground is wet, the game is canceled. It rained today, so the game is __. {**canceled**, continued} |
| Math | Second grade arithmetic exam: 7 × 6 = __. {**42**, 36} |
| Coding | def factorial(n): return 1 if n == 0 else n * __. {**factorial(n-1)**, factorial(n+1)} |

# 1. Pre-Training

- Example: English-French translation
- The knowledge **naturally occurs** in **web-scale** corpora
- Uses **next-token prediction** to learn the knowledge from web-scale corpora

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,’” Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

[Radford et al. 2019]

# 1. Pre-Training

- Next-token prediction **doesn't need anything else** but corpora
- Less requirements means dealing with more data: ebook, code, html…

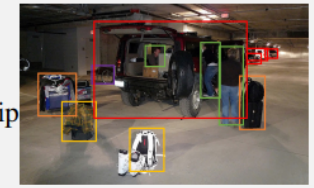
- No human annotations → **just raw text**
- No format requirement → **just raw text**
- No structure parsing → **just raw text**

- **Scalability!** → **Web-scale training data**

# 2. Supervised Fine-Tuning (SFT)

- Learns specified knowledge (to be an expert) from **carefully designed and annotated data** (supervised)
- The annotations demonstrate how human responds to an instruction
- Collecting annotations is expensive (by human and/or GPT)

**Context type 1: Captions**
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



**Context type 2: Boxes**
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

---

**Response type 1: conversation**
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
**Response type 2: detailed description**
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
**Response type 3: complex reasoning**
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

[Liu et al. NeurIPS'23]

# 2. Supervised Fine-Tuning (SFT)

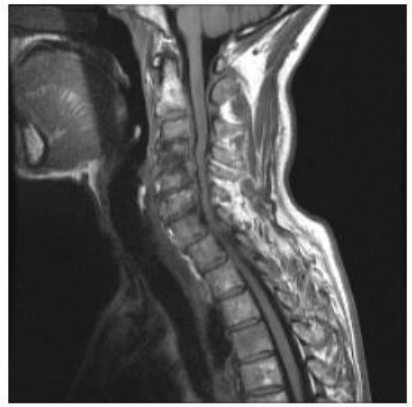
- Uses next-token prediction to learn as well
- 背解答模板

$$\text{Input} \rightarrow \underbrace{\text{Step}_1 \rightarrow \text{Step}_2 \rightarrow \cdots \rightarrow \text{Step}_n}_{\text{Reasoning chain}} \rightarrow \text{Answer}$$

SFT data

**Question:** What is the finding in this image?

A) Stomach ulcer, B) Nerve compression,
C) Cord pathology, D) Liver cirrhosis



**Answer:** To determine the correct answer, we need to identify the finding in the image provided. The MRI scan in question is likely a view of a spine, not the stomach, or the liver. Stomach ulcers, liver cirrhosis, and spinal cord pathology can be diagnosed with MRI, but the image in question shows a spinal view, not a gastrointestinal or abdominal view. Therefore, the only reasonable options left are nerve compression and cord pathology.

C

[Pan et al. MICCAI'25]

# 3. Reinforcement Learning (RL)

學會自己推理
融會貫通

- Learns to reason by itself → better generalization
- The RL reward evaluates the quality of the final answer while leaving the reasoning process unconstrained

$$\boxed{\text{Input}} \rightarrow \underbrace{\text{Step}_1 \rightarrow \text{Step}_2 \rightarrow \cdots \rightarrow \text{Step}_n}_{\text{Reasoning chain}} \rightarrow \boxed{\text{Answer}}$$

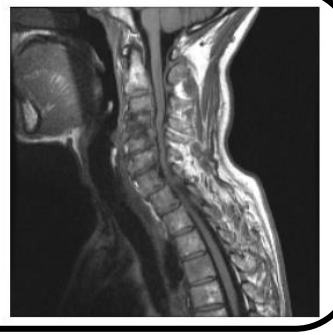
RL data

- RL training data require only the final answer to be annotated → much less expensive than SFT data

**Question:** What is the finding in this image?

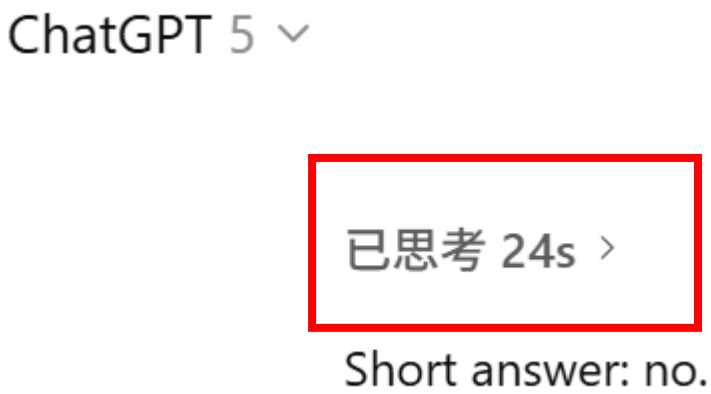
A) Stomach ulcer, B) Nerve compression,
C) Cord pathology, D) Liver cirrhosis

**Answer:** C



[Pan et al. MICCAI'25]

# 3. Reinforcement Learning (RL)

# Comparison

| Stage | Goal | Data | Data cost | Data amount | Learning signal |
|---|---|---|---|---|---|
| **1. Pre-Training** | Learns language modeling and generic knowledge | Massive unlabeled corpus | Low | Web-scale | Next-token prediction |
| **2. SFT** | Learns specified knowledge to be an expert | High-quality labeled instructions | Very high | Thousands to millions | Next-token prediction |
| **3. RL** | Learns to reason by itself for better generalization | Labeled final answers | Medium | Thousands to millions | Reward score |

# So, How to Train Our Own Expert LLM?

- Pre-training is crazy expensive 
- We can only handle the cost of **post-training** ☺

1. Starts from an open-source pre-trained model (Qwen, DeepSeek, etc.)
2. Collects and annotates our own SFT data and RL data for our tasks
3. Performs SFT and RL using our training data

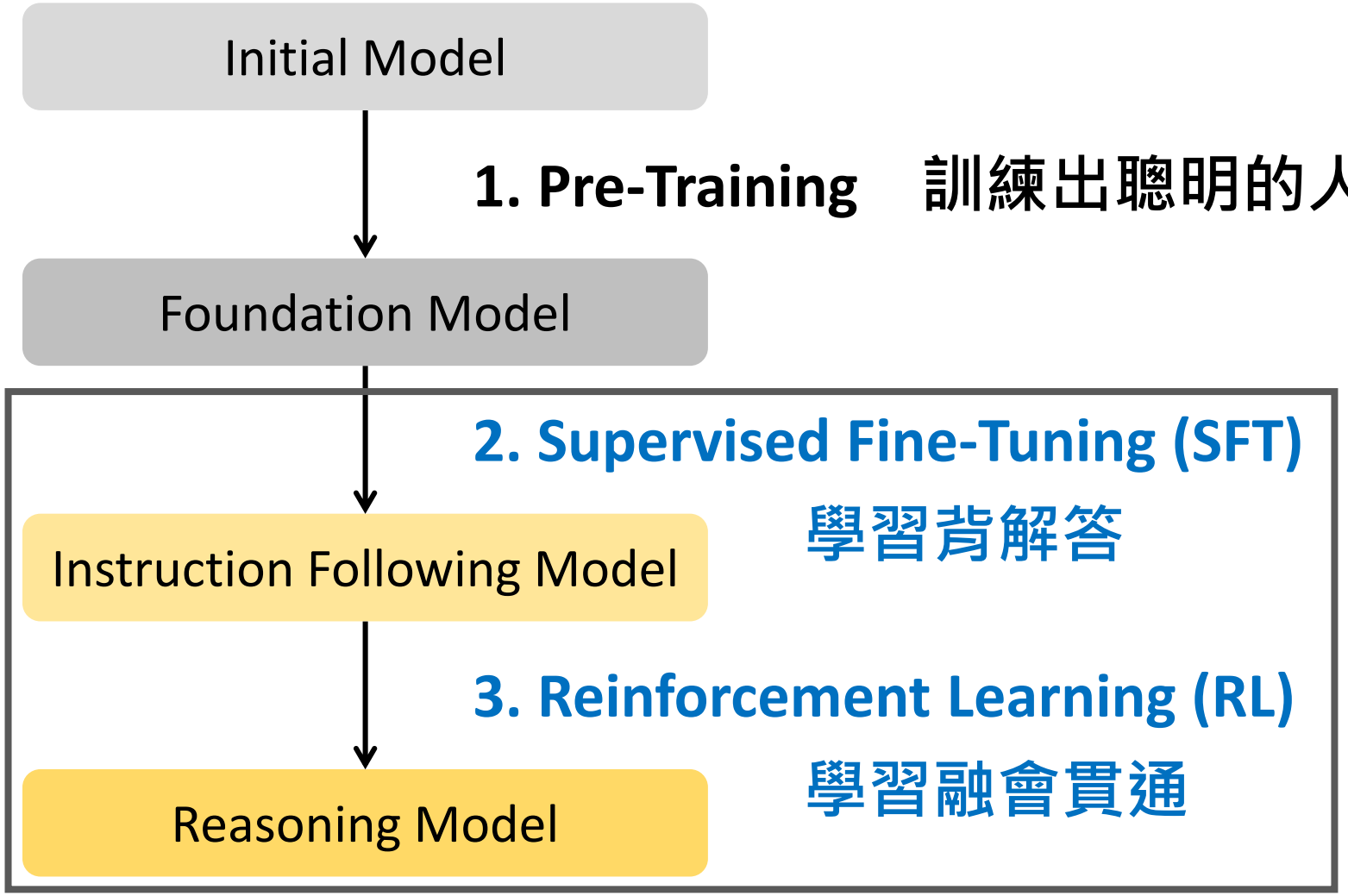
$$\text{Input} \rightarrow \underbrace{\text{Step}_1 \rightarrow \text{Step}_2 \rightarrow \cdots \rightarrow \text{Step}_n}_{\text{Reasoning chain}} \rightarrow \text{Answer}$$

SFT data

$$\boxed{\text{Input}} \rightarrow \underbrace{\text{Step}_1 \rightarrow \text{Step}_2 \rightarrow \cdots \rightarrow \text{Step}_n}_{\text{Reasoning chain}} \rightarrow \boxed{\text{Answer}}$$

RL data

# Summary




**Post-Training**

**將人才訓練成專家**