

Single Answer is Not Enough

On Generating Ranked Lists with
Medical Reasoning Models

Pittawat Taveekitworachai

Research Scientist

Typhoon (R&D), SCB 10X

What Is Typhoon?

Typhoon is an **advanced research initiative** focused on developing **open-source language technologies for the Thai language**. We provide **models, datasets, tools, and research** to advance Thai language AI and multimodal capabilities



Efficient Speed & Cost



Improved Thai Knowledge
and Instruction-Following
Performance



Open Source

Open access to resources fosters collaboration and drives AI innovation

A Collaboration Between



SiData+ Conference 2026 Project Members



Pittawat Taveekitworachai
Research Scientist



Krittaphas Chaisutyakorn
Clinical Ambassador



Natpatchara Pongjirapat
Clinical Ambassador



Kunat Pipatanakul
Lead AI Scientist



Piyalitt Ittichaiwong
Clinical Ambassador



Tossaporn Saengja
Data Scientist

Large Language Models → Reasoning Models



System 1 Thinking
Fast, common problems
Response immediately



System 2 Thinking
Slow, problem solving
Thinking longer for better answers

Large Language Models → Reasoning Models



LLM



Reasoning
Model

System 1 Thinking

Fast, common problems
Response immediately

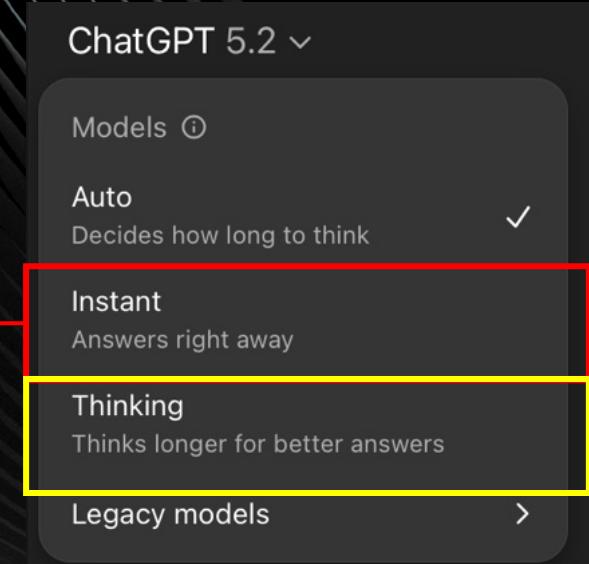
System 2 Thinking

Slow, problem solving
Thinking longer for better answers

Large Language Models → Reasoning Models



LLM



Reasoning Model

System 1 Thinking
Fast, common problems
Response immediately

System 2 Thinking
Slow, problem solving
Thinking longer for better answers

Large Language Models → Reasoning Models

ChatGPT 5.2 Instant

What is Typhoon AI?

Typhoon AI refers to a family of **large language models (LLMs)** and related AI tools developed primarily for the Thai language and local use cases. It isn't a weather phenomenon or unrelated malware group — in this context it's a Thai-focused artificial intelligence technology initiative. [Typhoon +1](#)

Core Description

- Typhoon is a series of **open-source large language models** optimized for Thai, developed to understand and generate Thai text more effectively than general global models. These models are intended to capture linguistic and cultural nuances specific to Thailand and improve natural language processing for Thai. [Typhoon +1](#)
- It is led by **SCB 10X** (the innovation arm of SCBX Group), which has released versions such as Typhoon-7B and later Typhoon 2, and provides them through APIs and repositories like Hugging Face. [Typhoon +1](#)

What It Can Do

- Natural language understanding and generation in Thai — the foundational capability for chatbots, virtual assistants, text summarization, translation, and more.

[OpenTyphoon Docs](#)

ChatGPT 5.2

Models ⓘ

Auto

Decides how long to think ✓

Instant

Answers right away

Thinking

Thinks longer for better answers

Legacy models >

What is Typhoon AI?

Thought for 8s >

"Typhoon AI" most commonly refers to the **SCB 10X** (the innovation arm of SCBX Group) and is positioned as a hub for open-source AI models (LLMs) that people can use to build Thai-capable AI systems. [Typhoon +2](#)

What you get under the Typhoon umbrella (examples):

- **Thai-focused LLMs** (e.g., "Typhoon 2.5" releases) intended for practical Thai workflows [Typhoon +1](#)
- Multiple modalities listed on the project site, including **text, reasoning, ASR (speech-to-text), OCR, and vision** [Typhoon](#)
- Availability via **open downloads** (at least some pretrained weights) and **API access** for certain instruction-tuned offerings [SCB](#)

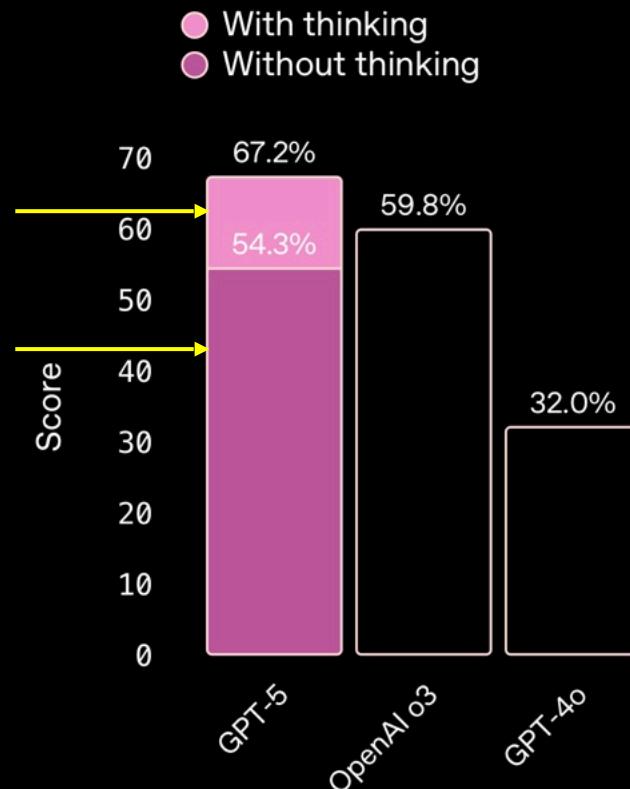
If you meant a different "Typhoon AI" (a company/product with that exact name), tell me where you saw it (link/app/context) and I'll pinpoint that specific one.

System 1 Thinking
Fast, common problems
Response immediately

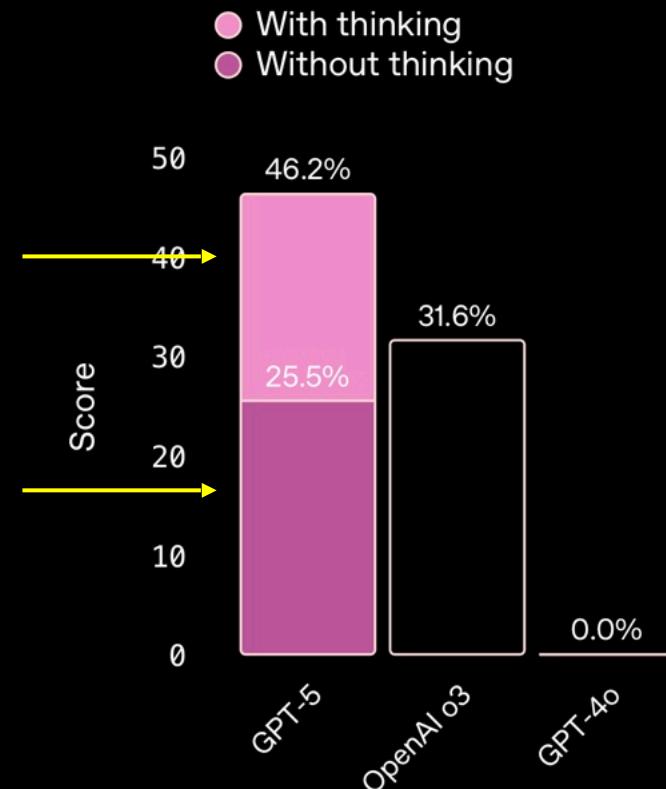
System 2 Thinking
Slow, problem solving
Thinking longer for better answers

Large Language Models → Reasoning Models

HealthBench
Realistic health conversations



HealthBench Hard
Challenging health conversations



Medical Reasoning Models



Reasoning Model

Medical Reasoning Models



Reasoning Model

+  Medical data

Specialized fine-tuning

Medical Reasoning Models



Reasoning Model

+  Medical data

Specialized fine-tuning



Medical Reasoning Model

Medical Reasoning Models (MRMs)



A patient presents with acute chest pain. What is the most likely diagnosis?

- A) Acute coronary syndrome, B) GERD, C) Costochondritis, D) Panic attack

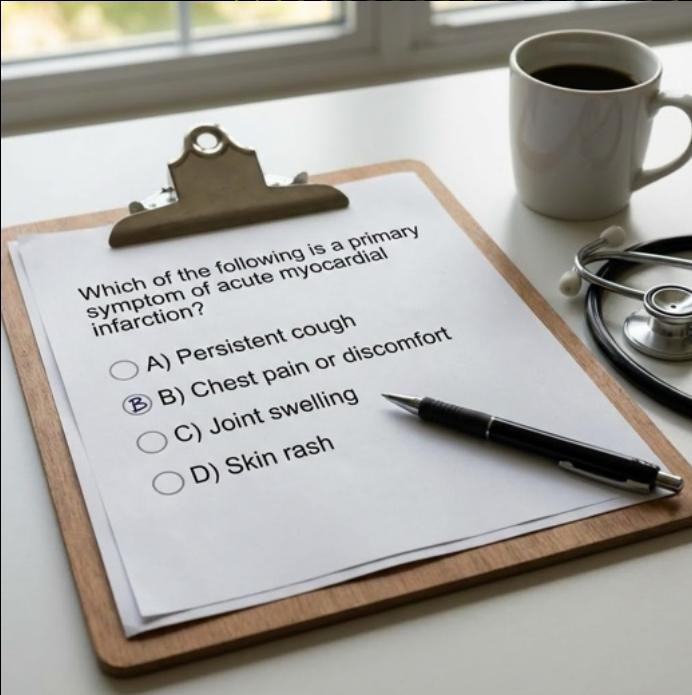


This is a multiple-choice question. In a clinical setting, when a patient presents with acute chest pain, the most critical and likely diagnosis to consider first is acute coronary syndrome (ACS) because it is life-threatening and requires immediate intervention. While GERD, costochondritis, and panic attacks are also possible causes of chest pain, the "most likely" diagnosis in the context of a medical emergency is often ACS until proven otherwise. However, we should note that the question does not provide any additional details about the patient's age, risk factors, or characteristics of the pain. In medical education, when a patient presents with acute chest pain, ...



- A) Acute coronary syndrome

However, Questions Are **Not Always Presented With Choices**



License exam



Case consultation



Differential diagnosis

However, Questions Are **Not Always Presented With Choices**

The Same Goes For **Medical Models**



MCQ

A patient presents with acute chest pain. What is the most likely diagnosis?

- A) Acute coronary syndrome, B) GERD, C) Costochondritis, D) Panic attack

A) Acute coronary syndrome



QA

A patient presents with acute chest pain. What is the most likely diagnosis?

Acute coronary syndrome



Ranked List

A patient presents with acute chest pain. What are the most likely diagnoses?

1. Acute coronary syndrome
2. Gastroesophageal reflux disease
3. Musculoskeletal chest wall pain
4. Pulmonary embolism
5. Aortic dissection

MRMs Do Not Always Follow Answer Format Instruction



General MRM

Thinking ...

Renal and bladder ultrasound



*"Please Answer As
A Ranked List"*

General MRM

Thinking ...



How **robustly** do MRMs follow **answer-format instructions** across different answer formats?



Answer-Format
Instruction



Medical
Reasoning Model



Ranked List



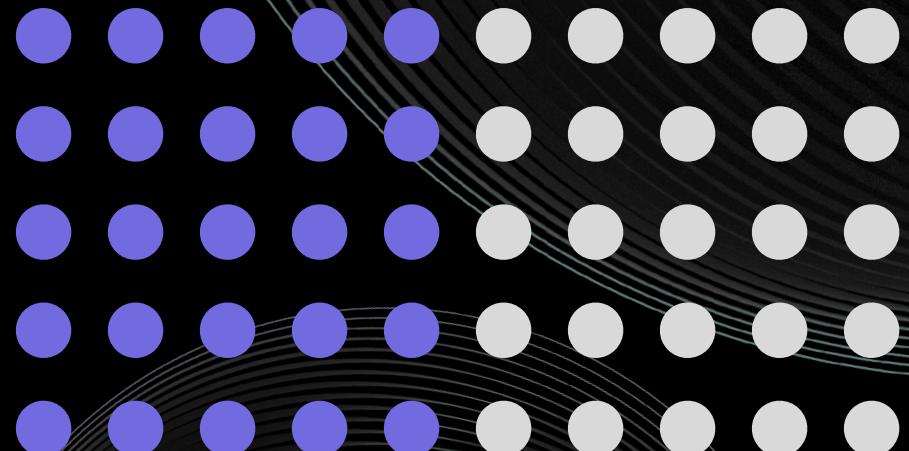
MCQ



QA

How Do We Measure Robustness?

$$Rbst_f(M, D) = \frac{1}{|D|} \sum_{q \in D} \text{Compiles}(M(q, f), f)$$



Robustness 50%

Half of the answers were successfully generated as a *valid ranked list*

"Please Answer As A Ranked List"

SiData+ Conference 2026 Models



2.5 Flash Lite / Flash / Pro

4.1 mini

2.5: 3B, 7B, 14B
3: 4B

4B

4B / 27B



MedGemma

Open Thoughts



OpenThinker 3:
7B

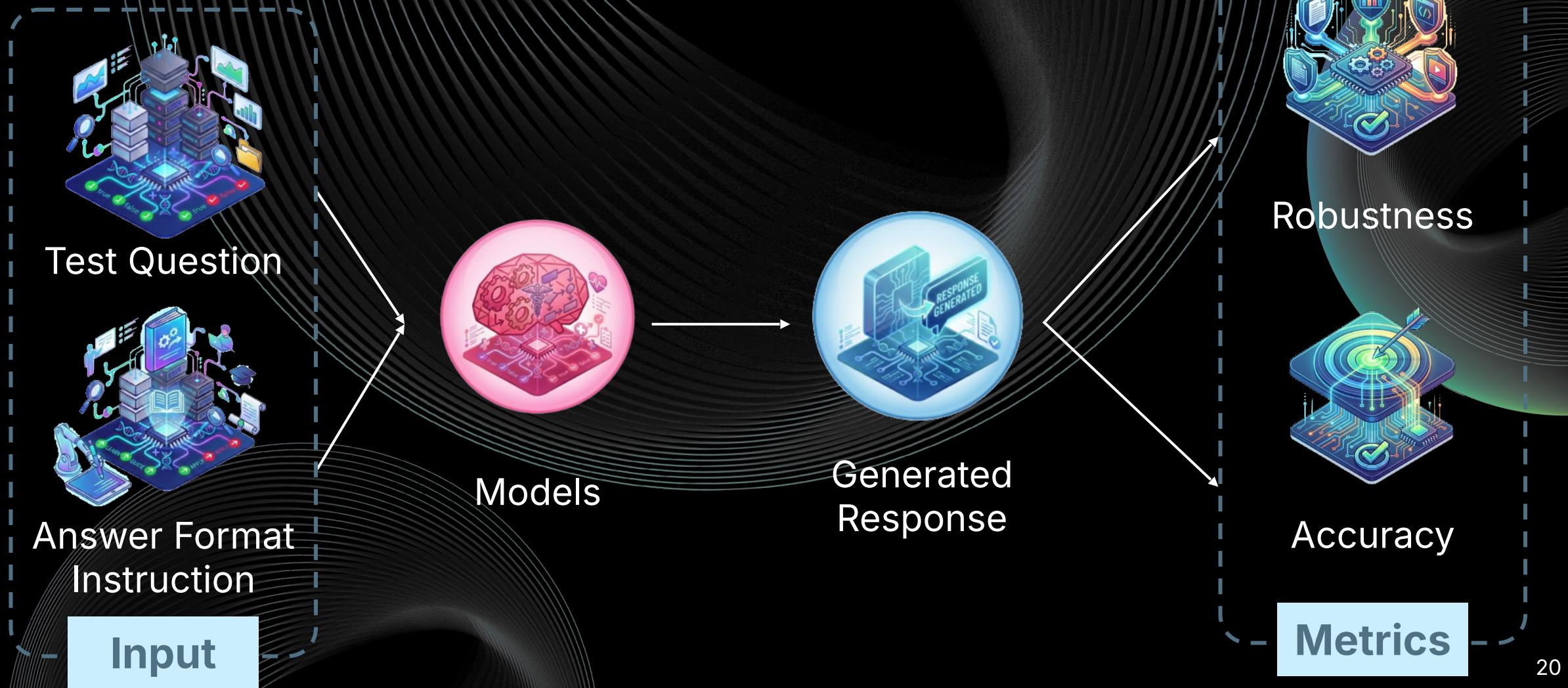
Medical Reasoning Models

1. HuatuoGPT o1 7B
2. m1 o1 7B
3. AlphaMed 7B



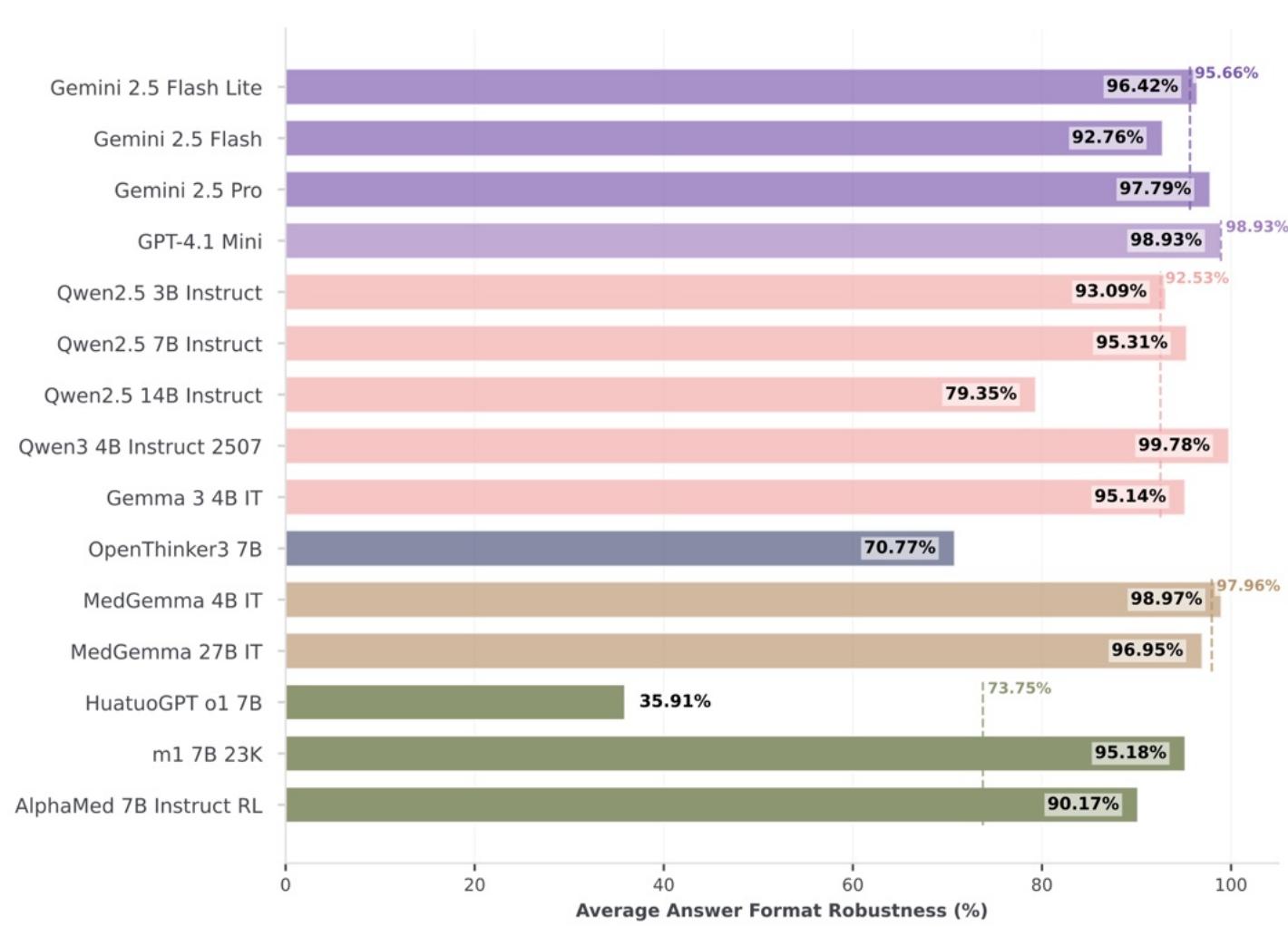
SiData+ Conference 2026

Experimental Setup



SiData+ Conference 2026

Findings

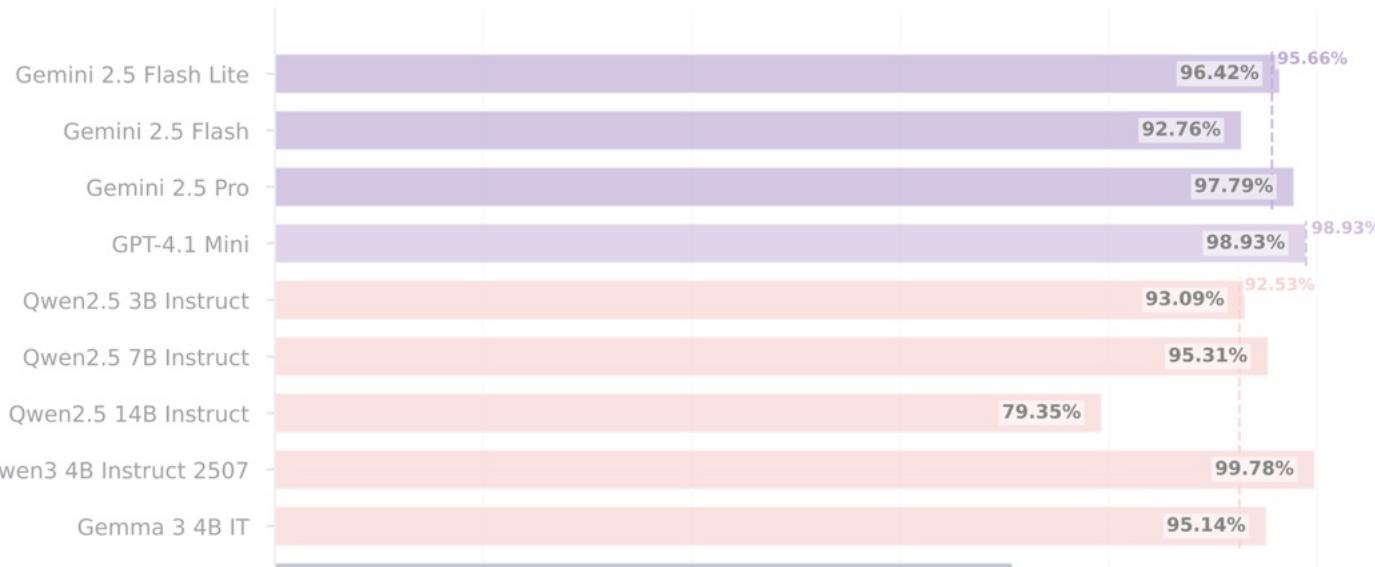


Finding 1

Most models are robust across different answer formats

SiData+ Conference 2026

Findings

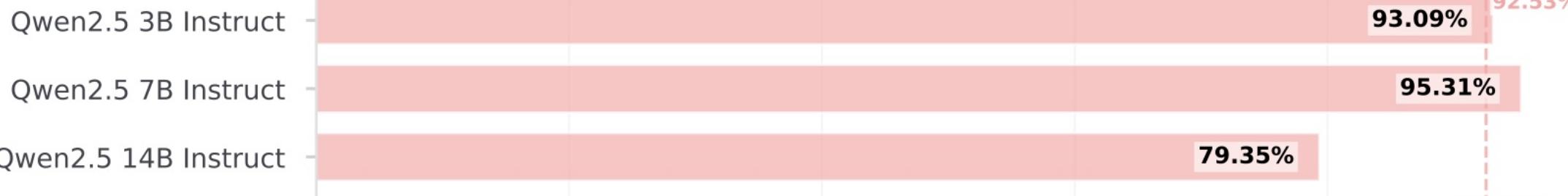


Finding 1

Most models are robust across different answer formats

Finding 2

Larger models are not necessarily better than smaller models

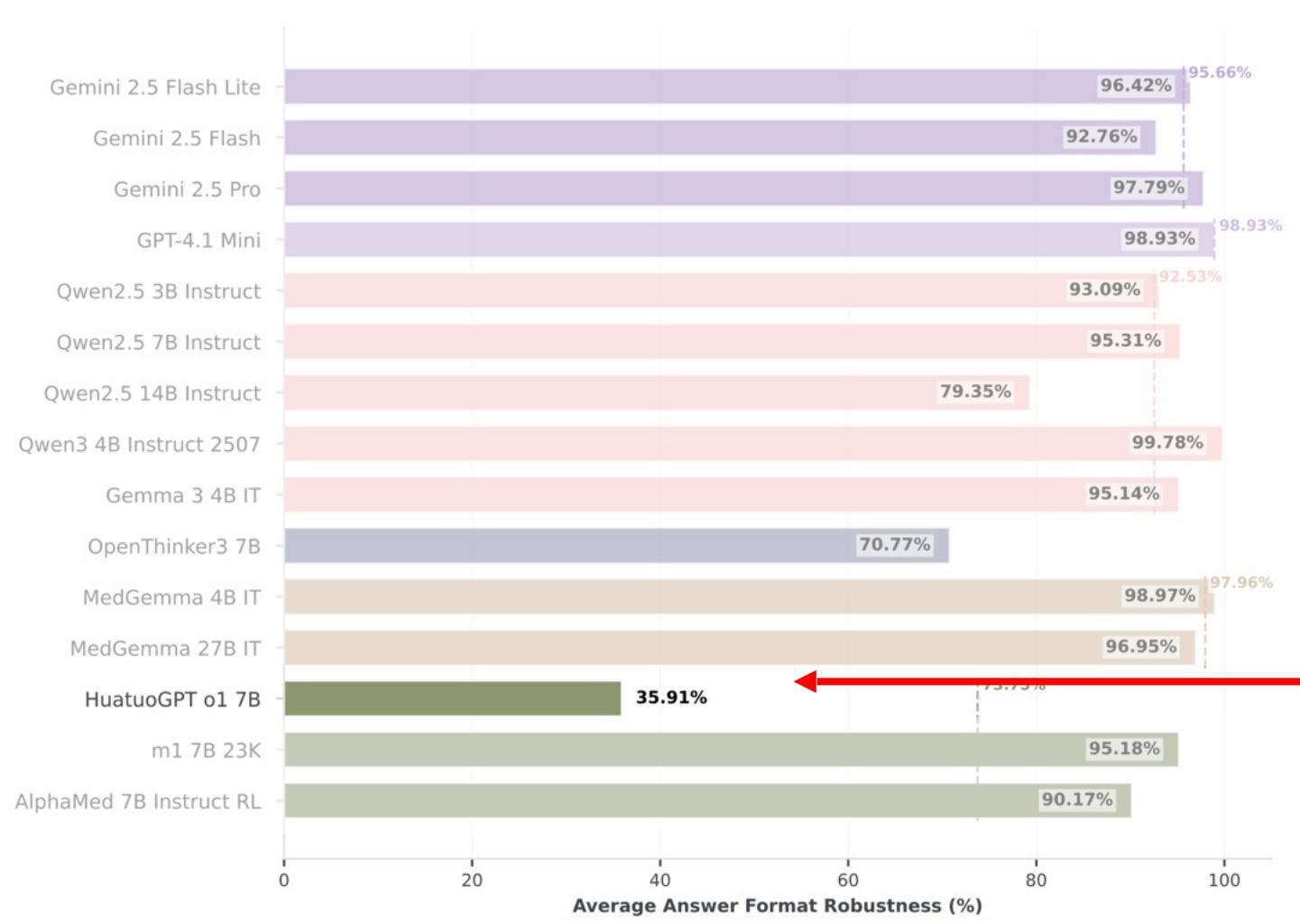


Alphameric 7B Instruct 2507



SiData+ Conference 2026

Findings



Finding 1

Most models are robust across different answer formats

Finding 2

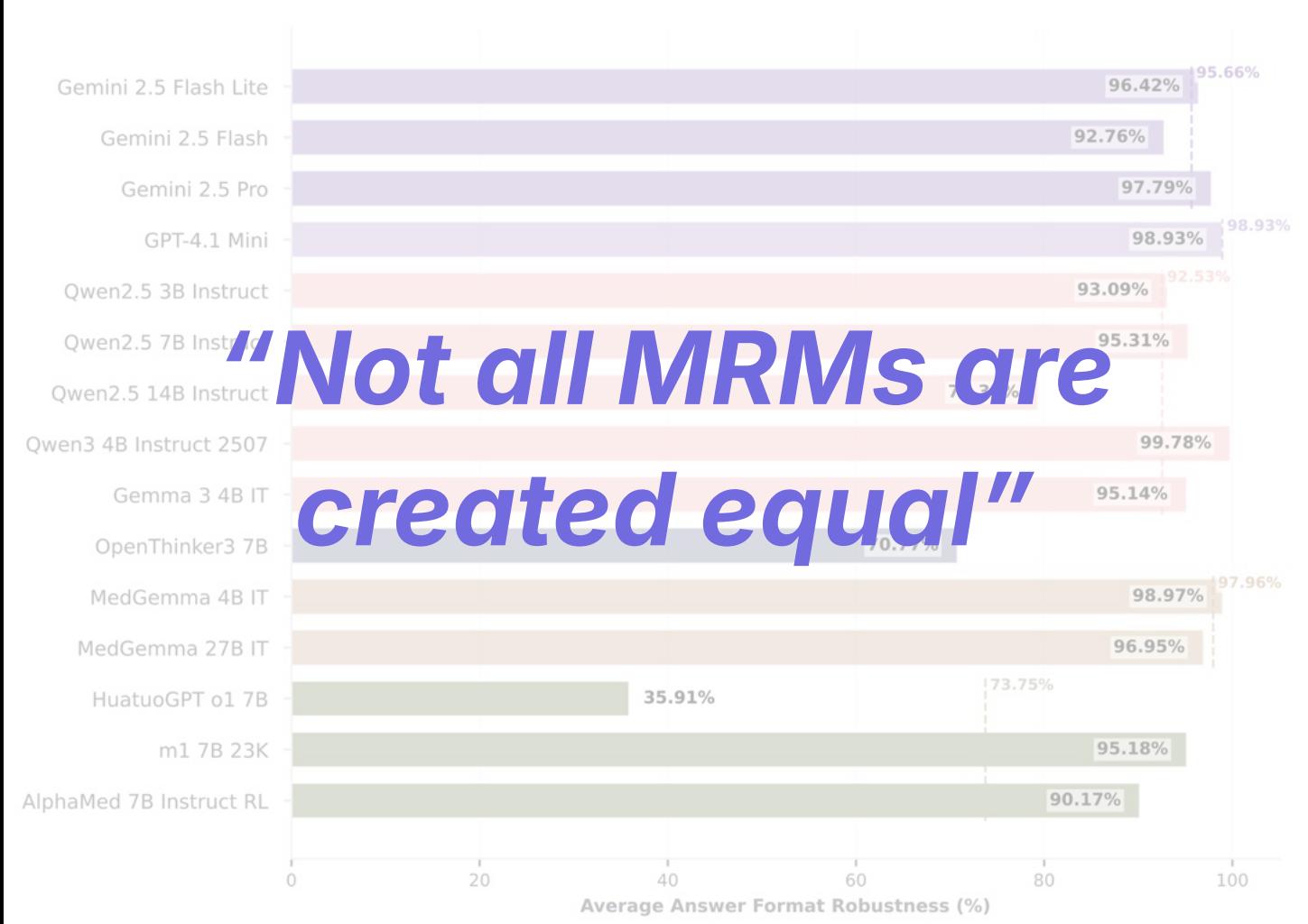
Larger models are not necessarily better than smaller models

Finding 3

HuatuoGPT has the worst robustness, unlike other MRMs

SiData+ Conference 2026

Findings



Finding 1

Most models are robust across different answer formats

Finding 2

Larger models are not necessarily better than smaller models

Finding 3

HuatuoGPT has the worst robustness, unlike other MRMs

How to Make a Model Think Longer?

Teach It

Supervised Fine-Tuning



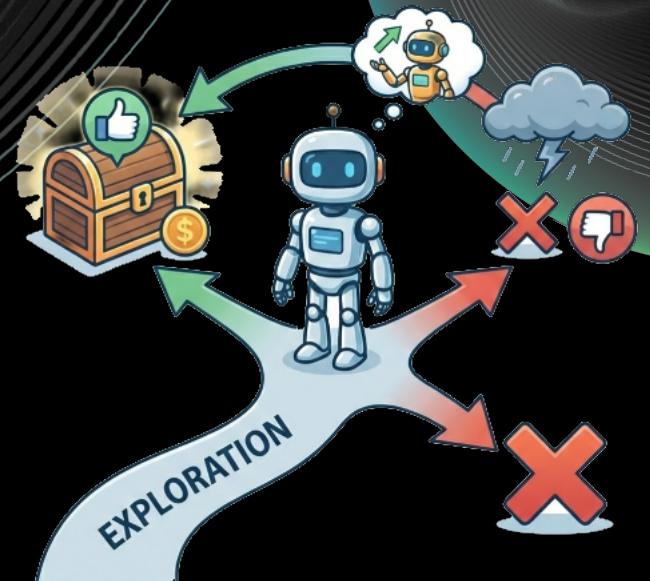
Ask It to Mimic

Knowledge Distillation



Let It Explore

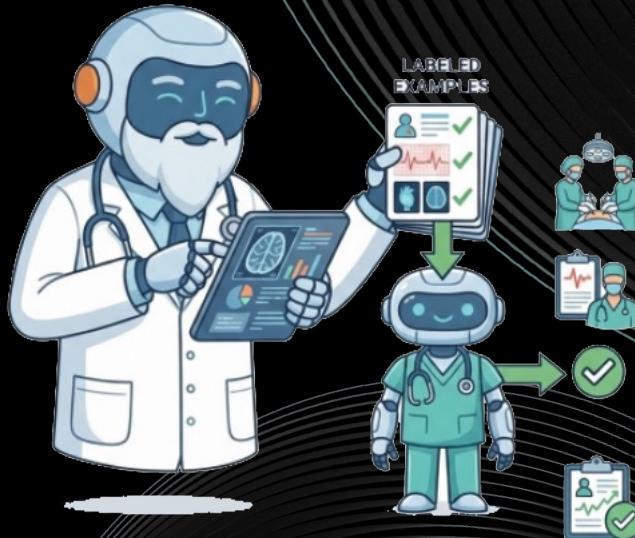
Reinforcement Learning



How to Make a Medical Reasoning Model?

Teach It

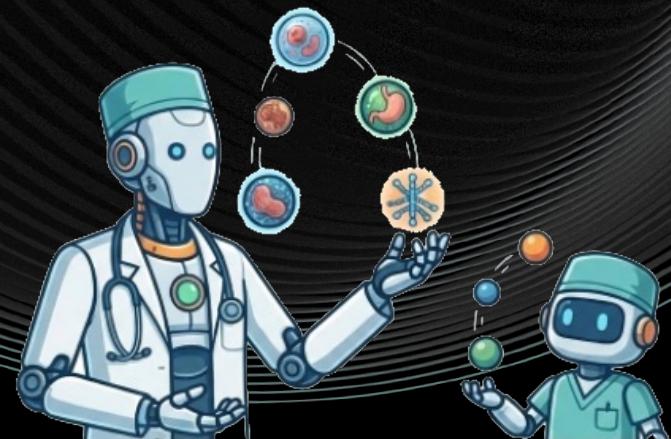
Teaching A Model Thinking Patterns



E.g., HuatuoGPT-01

Ask It to Mimic

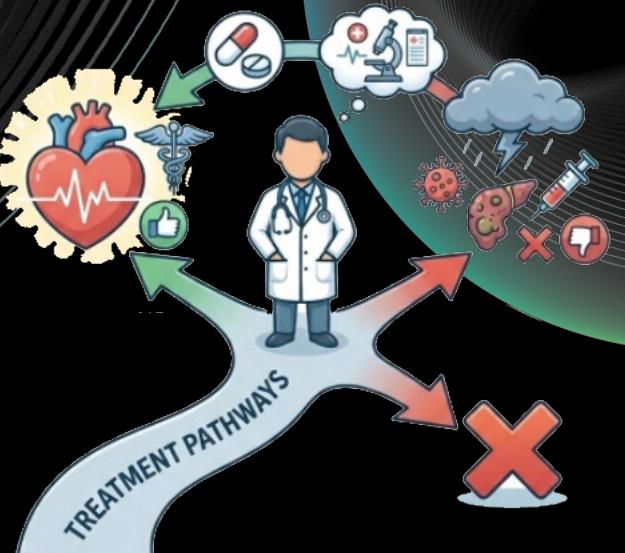
Imitating Teacher Thinking Patterns



E.g., m1

Let It Explore

Explore On Its Own How To Think



E.g., AlphaMed

Why Do We Need Controlled Experiments?

Problem

Comparing closed MRMs makes it very difficult to **identify training factors that affect robustness**



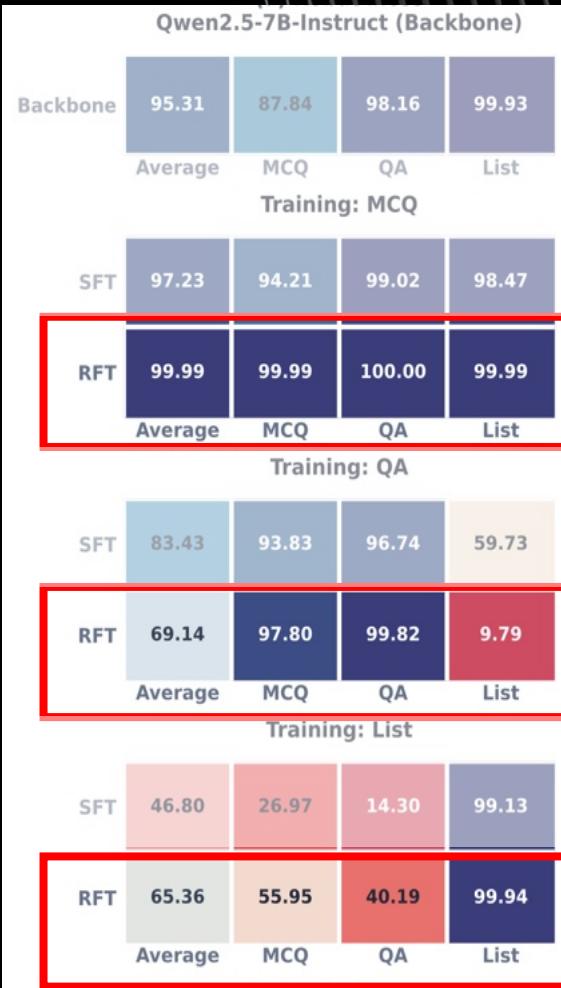
Solution

Controlled fine-tuning enables **better factor isolation** for causal analysis of MRMs



SiData+ Conference 2026

Findings



Finding 1

RFT yields very high cross-format robustness when trained on common formats (MCQ) but severe brittleness on rare formats (List/QA), whereas SFT degrades more moderately, making RFT both more powerful and more fragile under format mismatch

SiData+ Conference 2026

Findings



Finding 1

RFT yields very high cross-format robustness when trained on common formats (MCQ) but severe brittleness on rare formats (List/QA), whereas SFT degrades more moderately, making RFT both more powerful and more fragile under format mismatch

Finding 2

Fine-tuning makes models cross-format brittle when the target format is misaligned (e.g., QA), while stability is preserved mainly when fine-tuning matches dominant formats (e.g., MCQ)

How Robustly Do Medical Reasoning Models Follow Answer-Format Instructions?

An evaluation across answer formats, prompting, and fine-tuning paradigms

Answer Formats (Section 2)

MCQ

A patient presents with acute chest pain. What is the most likely diagnosis?
A) Acute coronary syndrome, B) GERD, C) Costochondritis, D) Panic attack

A) Acute coronary syndrome

QA

A patient presents with acute chest pain. What is the most likely diagnosis?

Acute coronary syndrome

Ranked List

A patient presents with acute chest pain. What are the most likely diagnoses?

1. Acute coronary syndrome
2. Gastroesophageal reflux disease
3. Musculoskeletal chest wall pain
4. Pulmonary embolism
5. Aortic dissection

Observational Analysis via Prompting (Section 4)



Answer-format
Instruction Following
(Finding 1)



Knowledge-Format
Entanglement
(Finding 2)



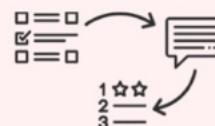
CoT Prompting
(Finding 3)

↓ Revisiting Observed Findings in More Controlled Settings ↓

Controlled Fine-Tuning Experiments (Section 5)



Training Paradigms
(SFT vs. RFT)
(Finding 4)



Cross-Format
Robustness
(Finding 5)



RFT Reward Design
and Robustness
(Finding 6)

Metrics (Section 3)



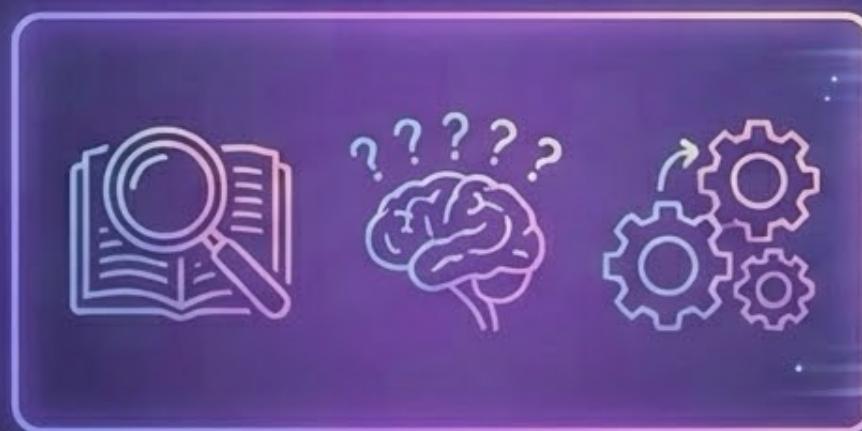
Correctness



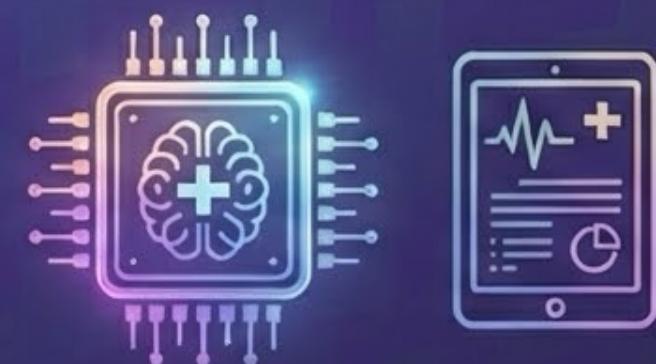
Robustness



List-Specific



Research



Medical Reasoning Model



Typhoon-SI
Med Thinking

Typhoon-Si Med Thinking 4B



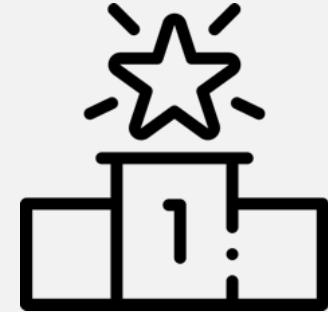
Reasoning Model

*Trained to think longer
for better results*



Small and Efficient

*Only at the 4B size, able to
fit in a single latest-
generation consumer GPU*



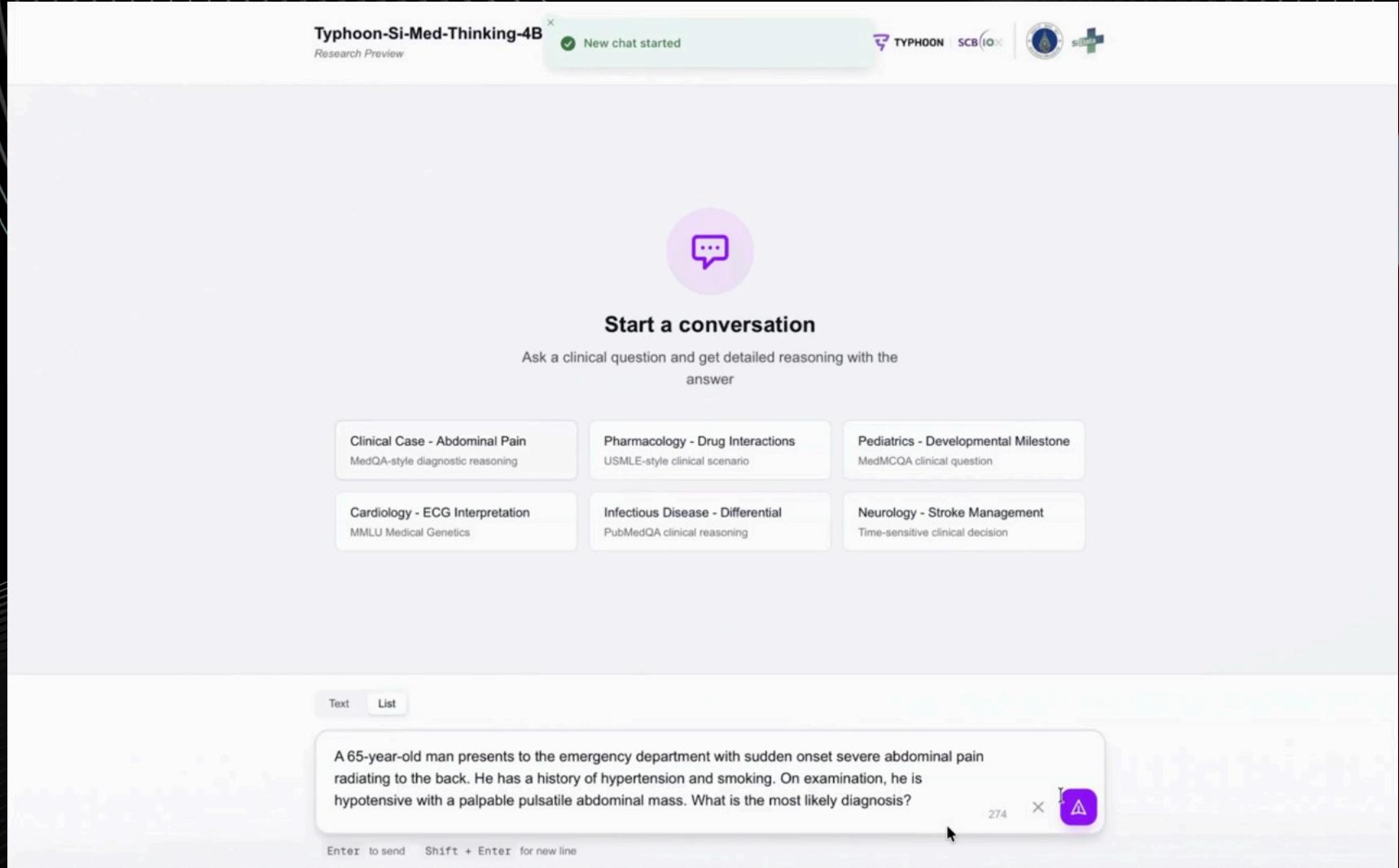
State-of-the-Art Performance

*Outperforms Gemini 2.5
Pro at ~100x lower cost*

Performance Results

Model Name	MCQ	QA	List (Acc)	List (MRR)
Gemini 2.5 Flash Lite	48.47	48.69	53.82	46.36
Gemini 2.5 Flash	<u>55.19</u>	46.10	62.66	53.52
Gemini 2.5 Pro	58.68	49.20	<u>68.46</u>	<u>58.85</u>
GPT-4.1-mini	54.72	47.02	61.71	53.82
MedGemma 4B it	37.09	43.19	53.34	38.65
MedGemma 27B it	48.97	47.64	50.74	43.33
Qwen3 4B Instruct 2507	43.82	47.22	53.01	40.49
Typhoon-Si-Med-Thinking-4B OURS	44.58	<u>48.87</u>	94.73	90.68

Demo: Ranked List Answer #1



The screenshot shows a web-based medical reasoning interface. At the top, a banner indicates a "New chat started" with "Typhoon-Si-Med-Thinking-4B" and "Research Preview". The interface features a large "Start a conversation" button with a speech bubble icon. Below it, a section titled "Ask a clinical question and get detailed reasoning with the answer" lists six clinical categories:

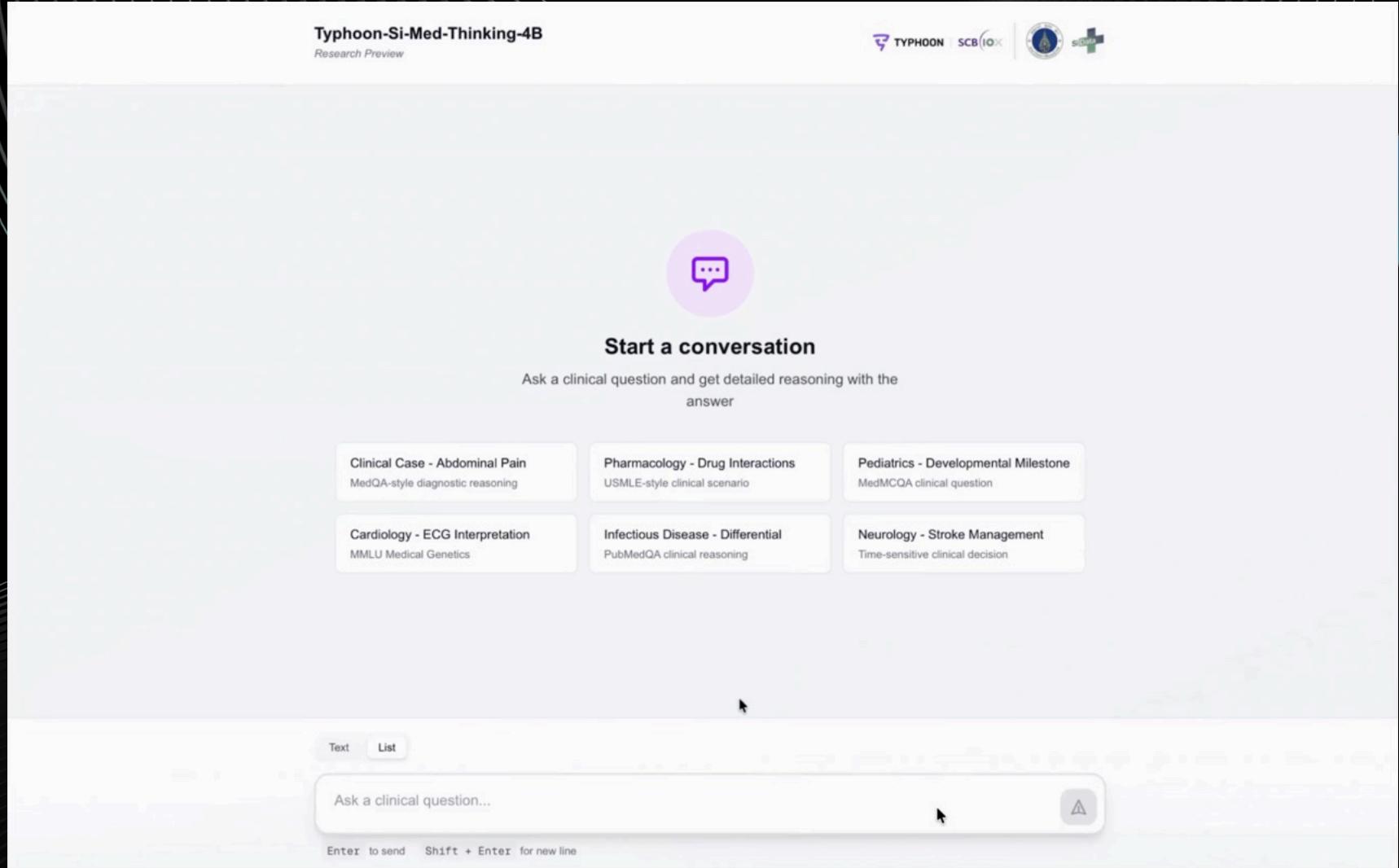
- Clinical Case - Abdominal Pain (MedQA-style diagnostic reasoning)
- Pharmacology - Drug Interactions (USMLE-style clinical scenario)
- Pediatrics - Developmental Milestone (MedMCQA clinical question)
- Cardiology - ECG Interpretation (MMLU Medical Genetics)
- Infectious Disease - Differential (PubMedQA clinical reasoning)
- Neurology - Stroke Management (Time-sensitive clinical decision)

At the bottom, a text input field contains the following clinical question:

A 65-year-old man presents to the emergency department with sudden onset severe abdominal pain radiating to the back. He has a history of hypertension and smoking. On examination, he is hypotensive with a palpable pulsatile abdominal mass. What is the most likely diagnosis?

Below the input field are buttons for "Text" and "List", and a "List" button is currently selected. The text input field has a character count of 274 and a delete button. The footer of the interface includes the text "Enter to send Shift + Enter for new line".

Demo: Ranked List Answer #2



The screenshot shows a user interface for a medical reasoning tool. At the top, it displays the title "Typhoon-Si-Med-Thinking-4B" and "Research Preview". To the right are logos for TYPHOON, SCBIOX, and SiData+.

In the center, there is a purple circular icon with a white speech bubble containing three dots, and the text "Start a conversation". Below this, a sub-instruction reads: "Ask a clinical question and get detailed reasoning with the answer".

Below the sub-instruction, there are six categories arranged in a 2x3 grid:

- Clinical Case - Abdominal Pain**
MedQA-style diagnostic reasoning
- Pharmacology - Drug Interactions**
USMLE-style clinical scenario
- Pediatrics - Developmental Milestone**
MedMCQA clinical question

- Cardiology - ECG Interpretation**
MMLU Medical Genetics
- Infectious Disease - Differential**
PubMedQA clinical reasoning
- Neurology - Stroke Management**
Time-sensitive clinical decision

At the bottom, there is a text input field with the placeholder "Ask a clinical question..." and a send button with a triangle icon. Below the input field, a note says "Enter to send Shift + Enter for new line".

Open Knowledge & Open Model

Preprint Paper



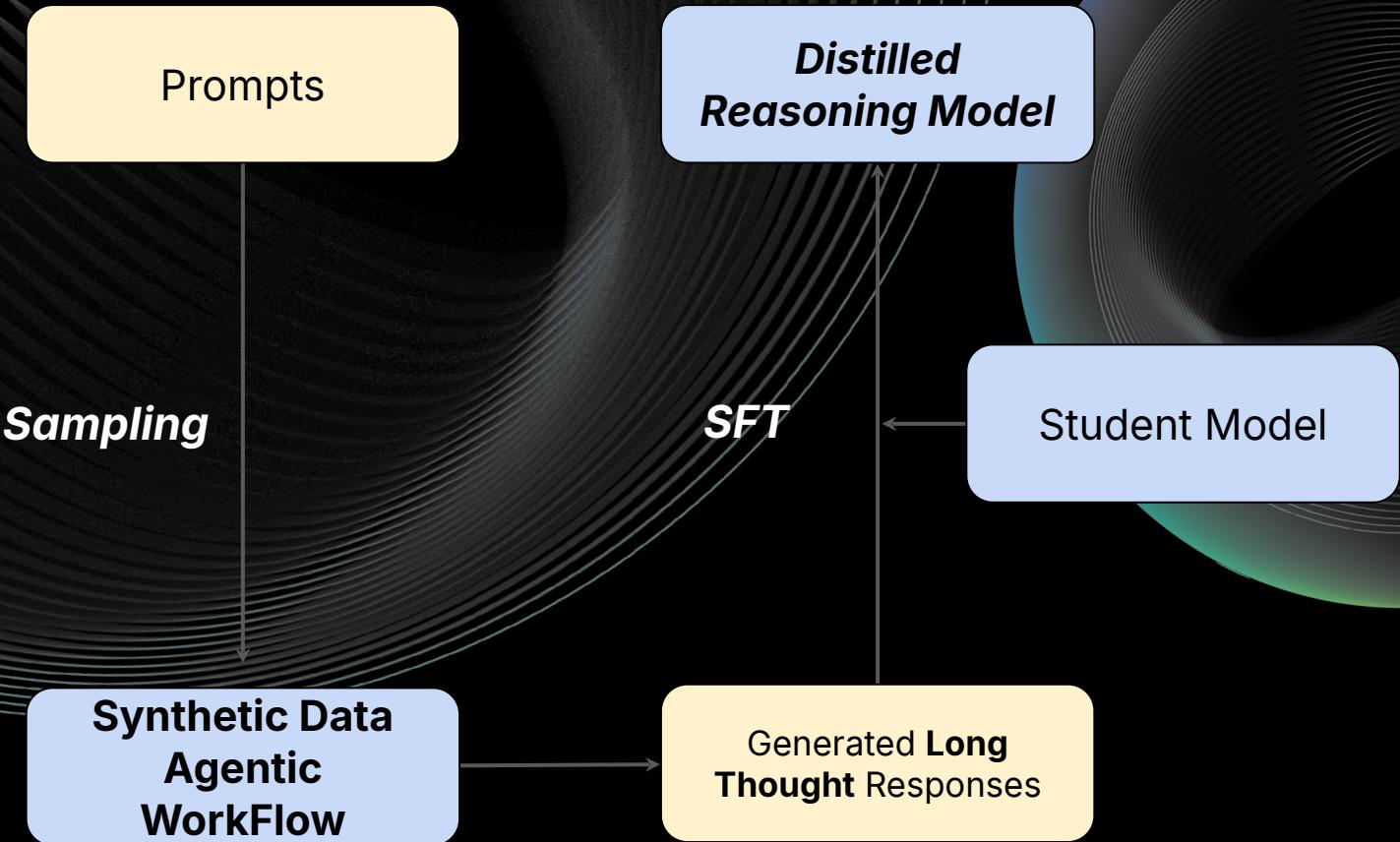
Models Are Available
On  Hugging Face



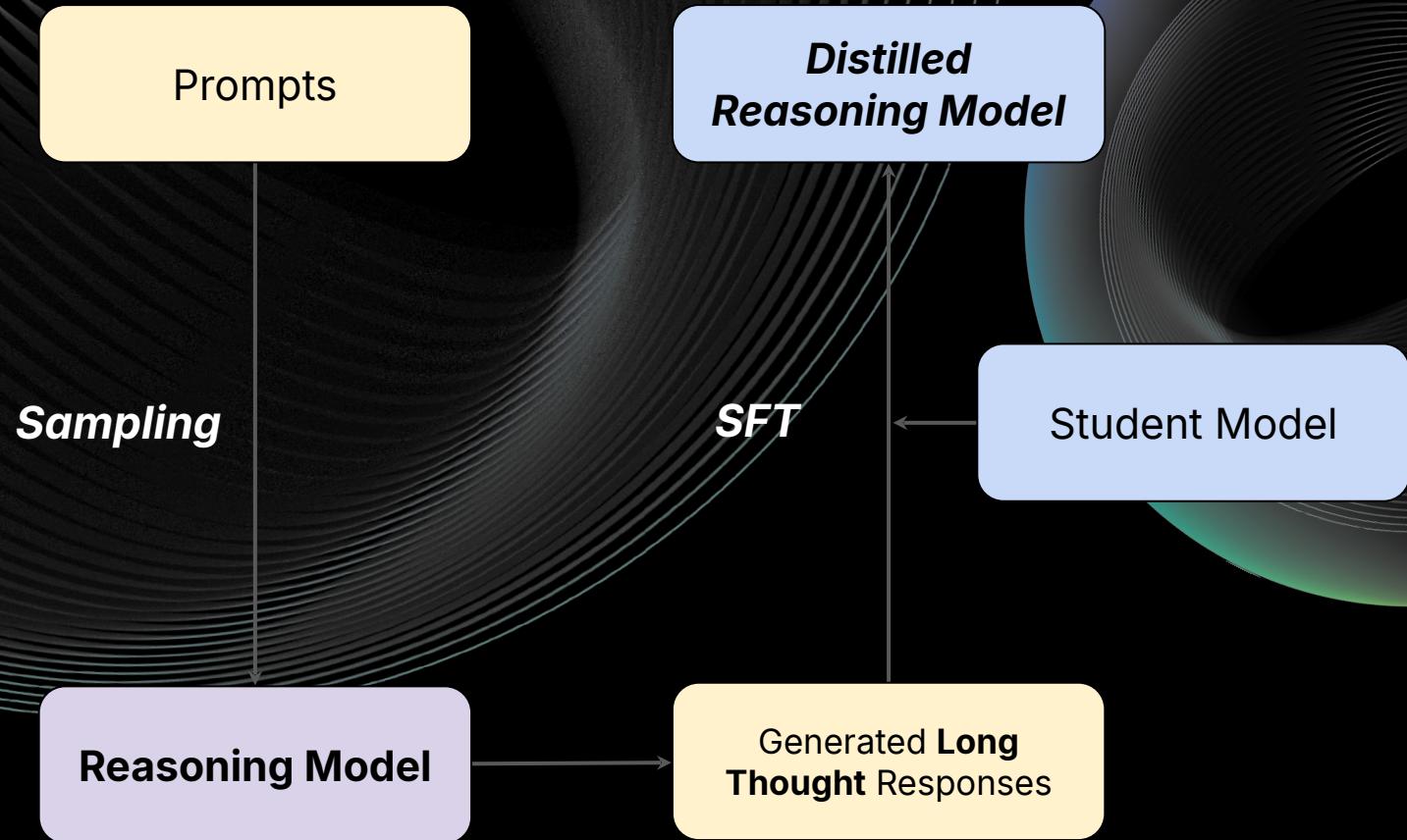
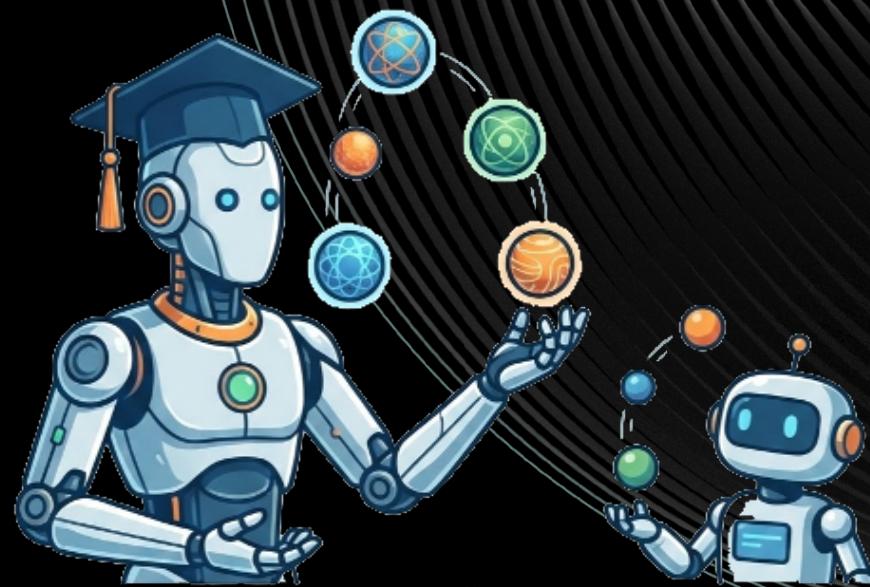
SiData+ nference 2026



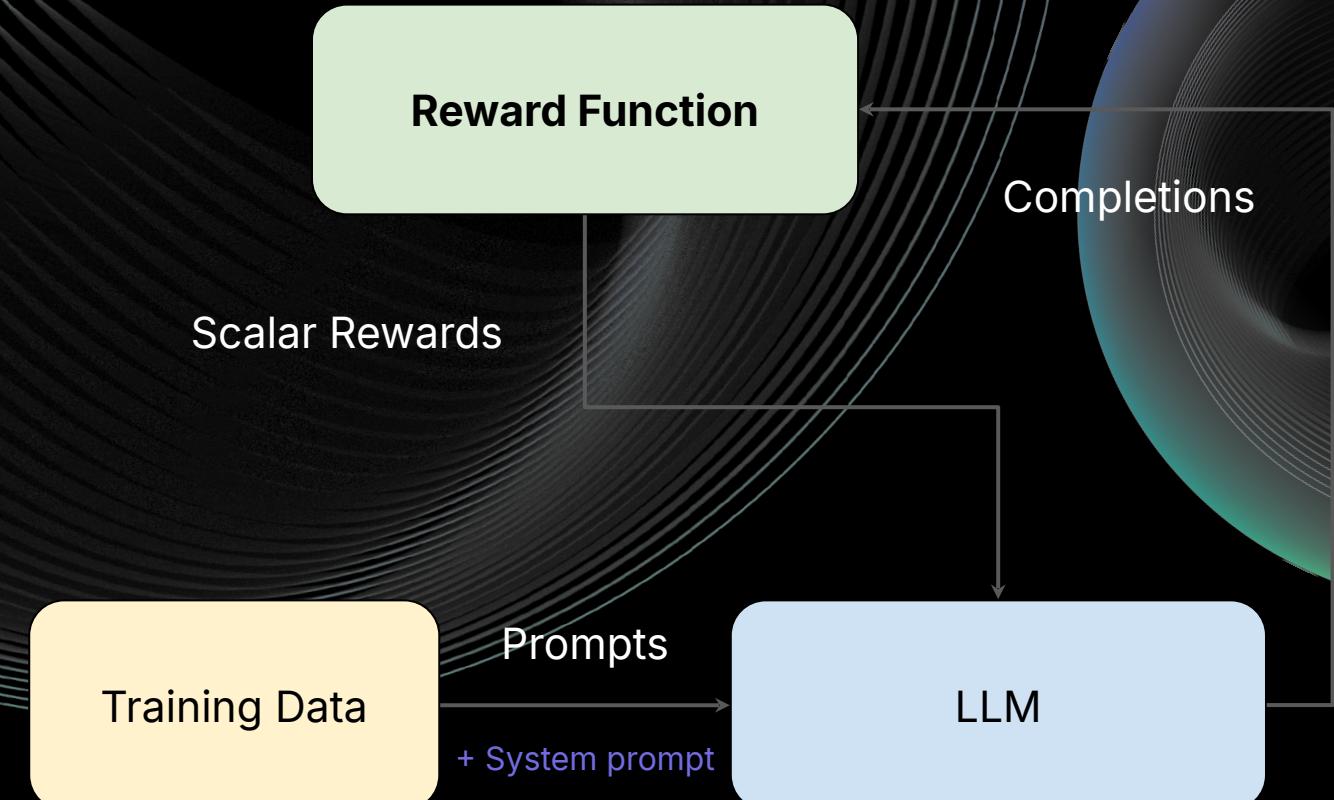
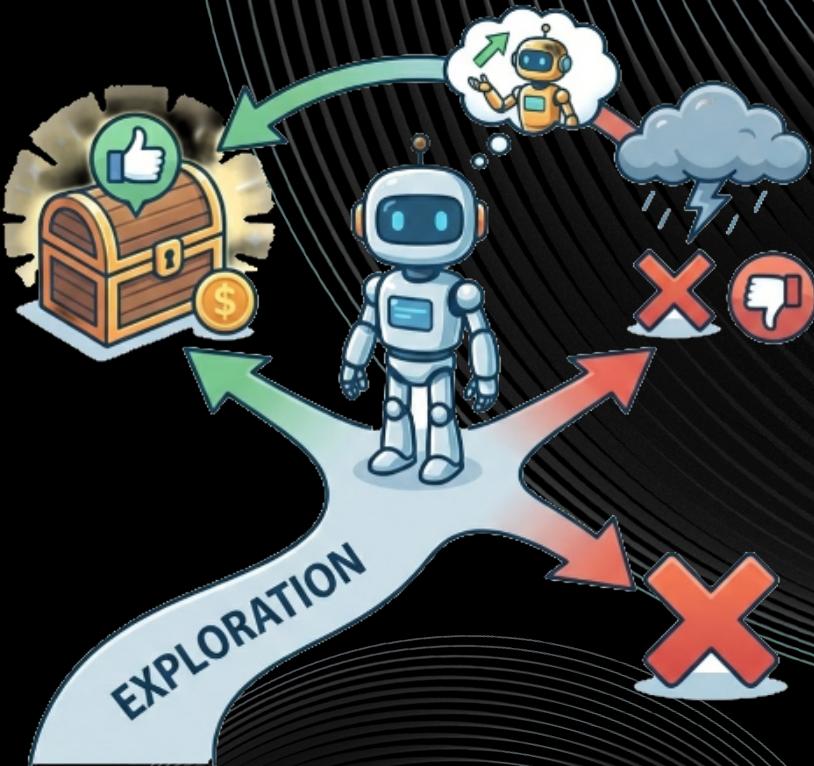
Supervised Fine-Tuning with Synthetic Data



Knowledge Distillation



Reinforcement Fine-Tuning



[...] The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`. [...]