SiData+ Conference 2026

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

SiData+
Conference 2026

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

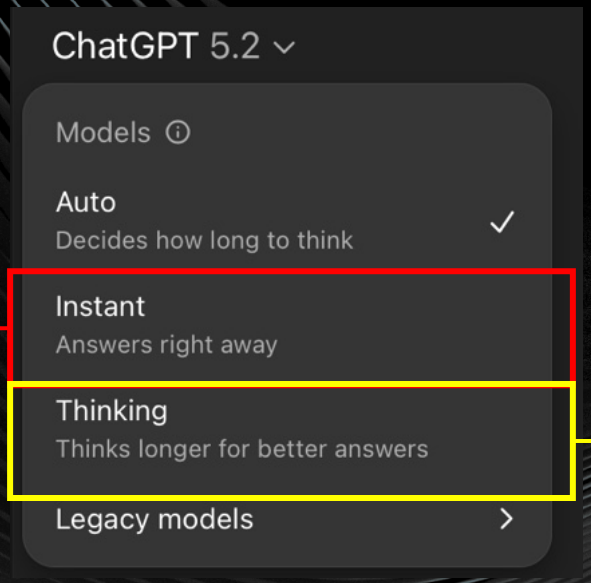
Reasoning Model

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



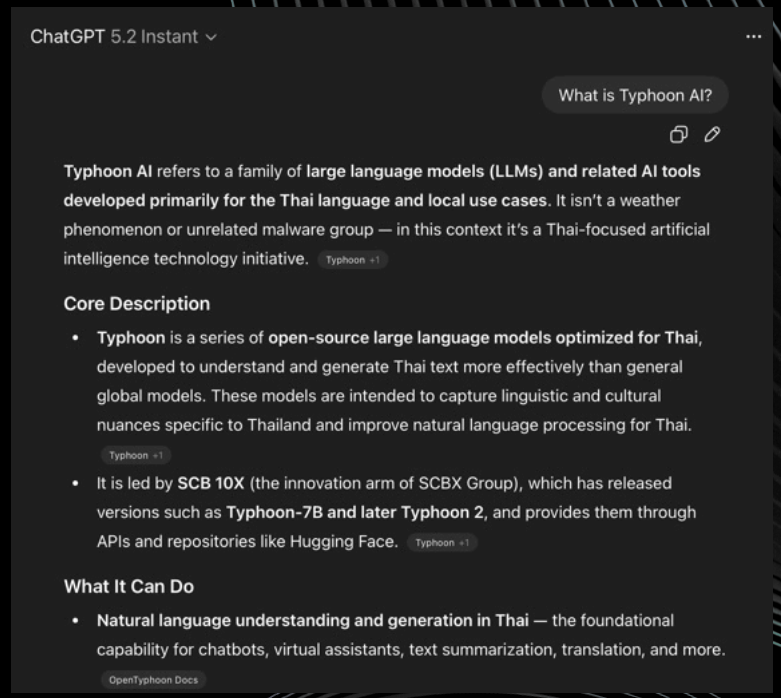
Reasoning Model

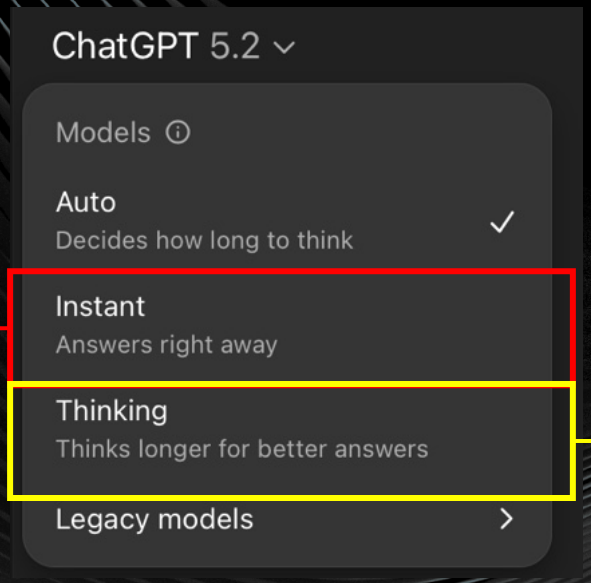
System 1 Thinking
Fast, common problems
Response immediately

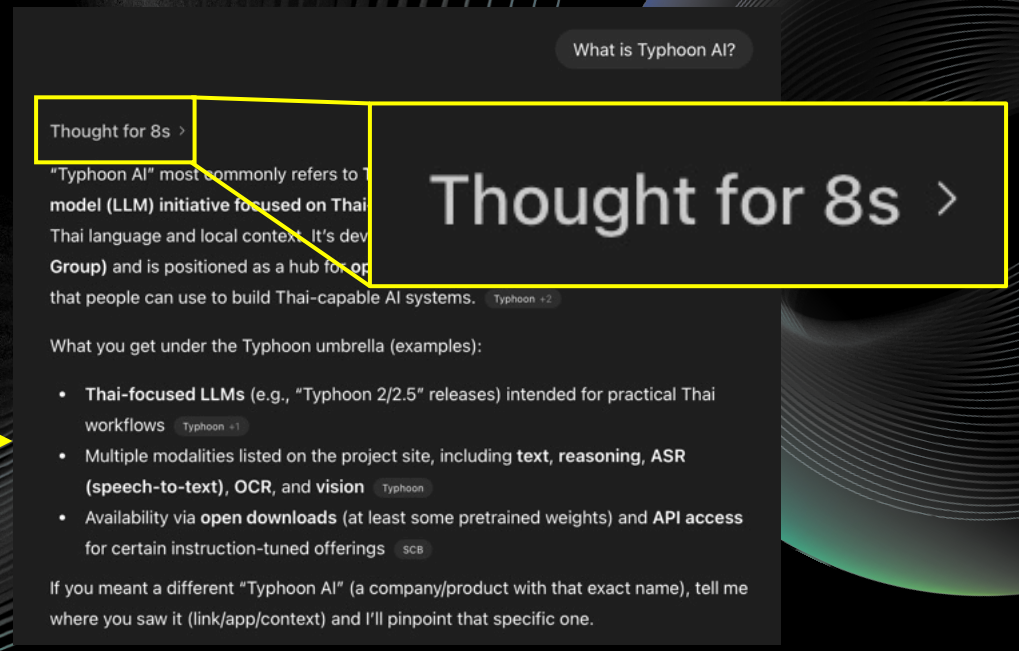
System 2 Thinking
Slow, problem solving
Thinking longer for better answers

SiData+
Conference 2026

Large Language Models → Reasoning Models





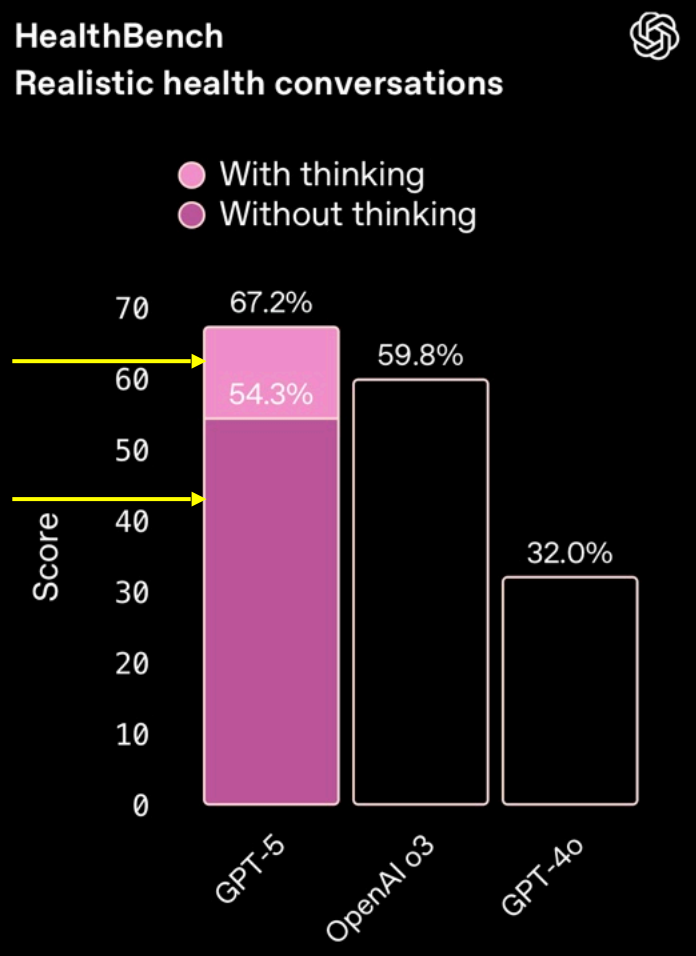


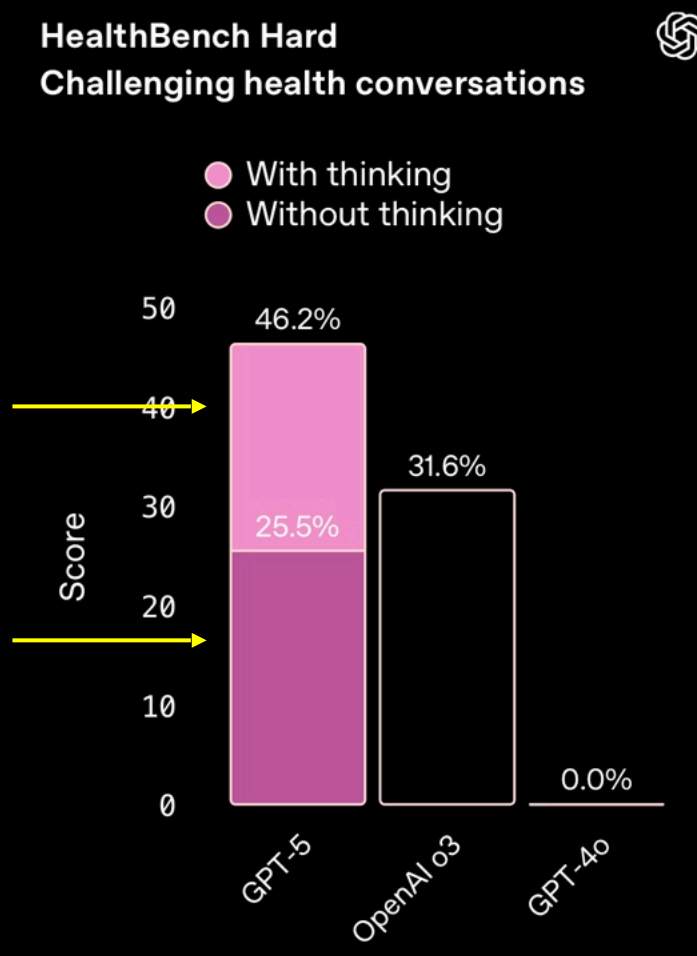
System 1 Thinking
Fast, common problems
Response immediately

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

SiData+
Conference 2026

Large Language Models → Reasoning Models





SiData+
Conference 2026

Medical Reasoning Models

Reasoning Model

SiData+
Conference 2026

Medical Reasoning Models

Reasoning Model

+ 💊 Medical data

Specialized fine-tuning

SiData+
Conference 2026

Medical Reasoning Models

Reasoning Model

+ 💊 Medical data

Specialized fine-tuning

Medical Reasoning Model

SiData+
Conference 2026

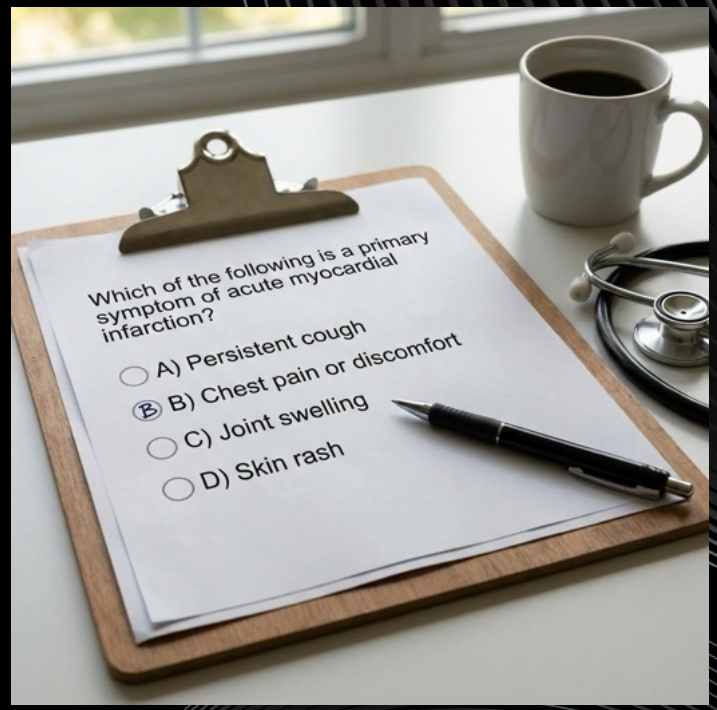
Medical Reasoning Models (MRMs)

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”

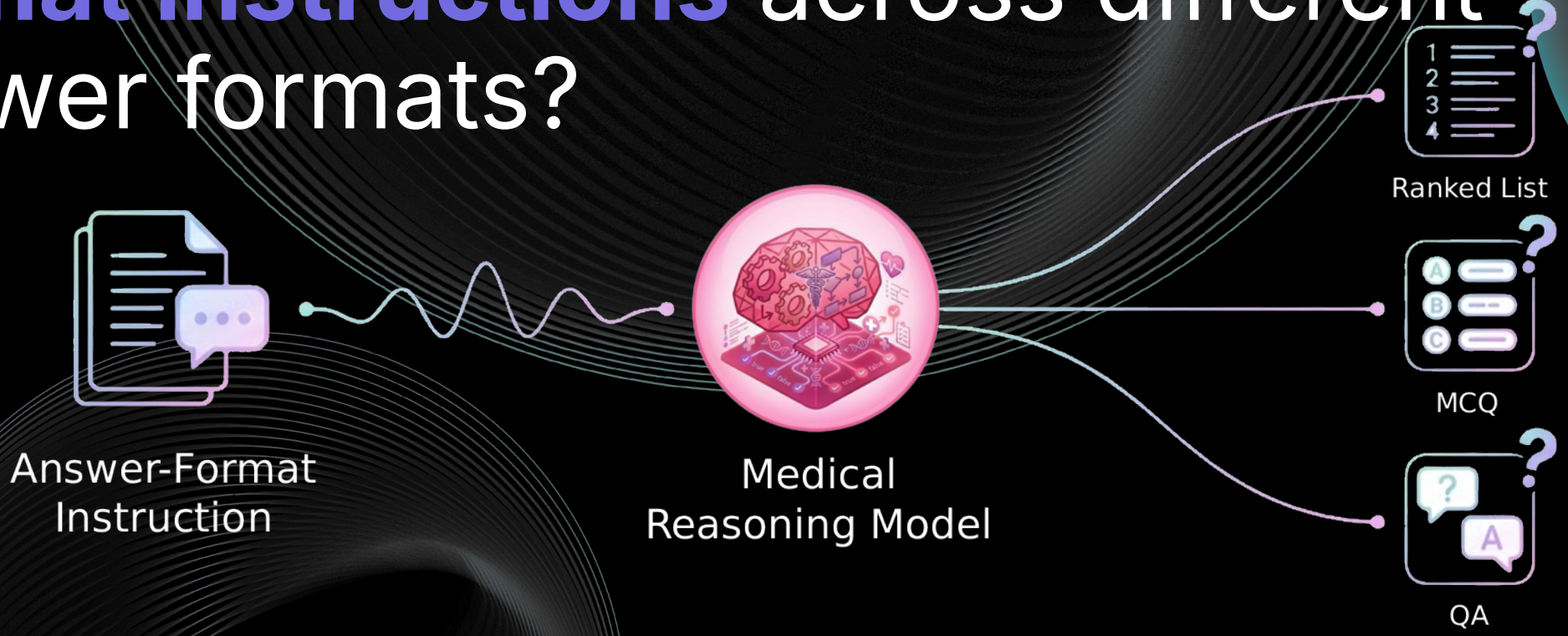
General MRM

💭 Thinking ...

SiData+
Conference 2026

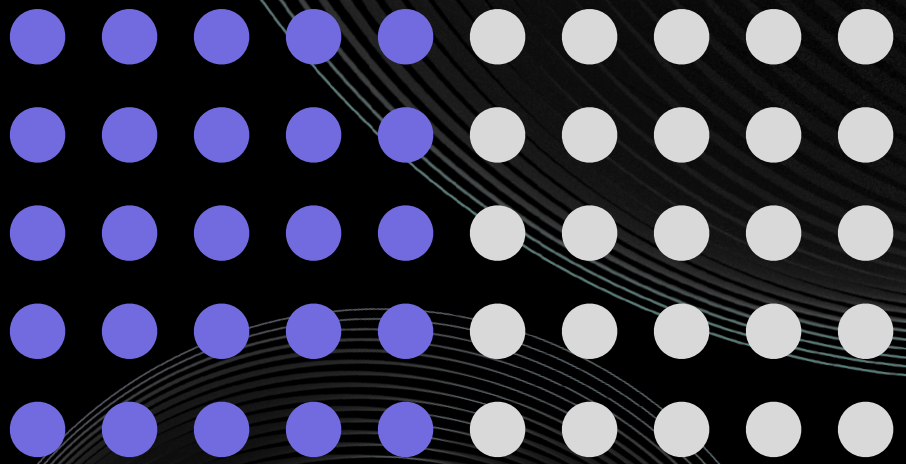
Research Question

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



How Do We Measure Robustness?

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



"Please Answer As *A Ranked List*"

Robustness 50%

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

SiData+
Conference 2026

Models

Gemini **2.5** Flash Lite / Flash / Pro

4.1 mini

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

Gemma 4B

MedGemma 4B / 27B

Open Thoughts **OpenThinker 3**: 7B

Medical Reasoning Models

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

SiData+
Conference 2026

Experimental Setup

Test Question

Answer Format Instruction

Input

Models

Generated Response

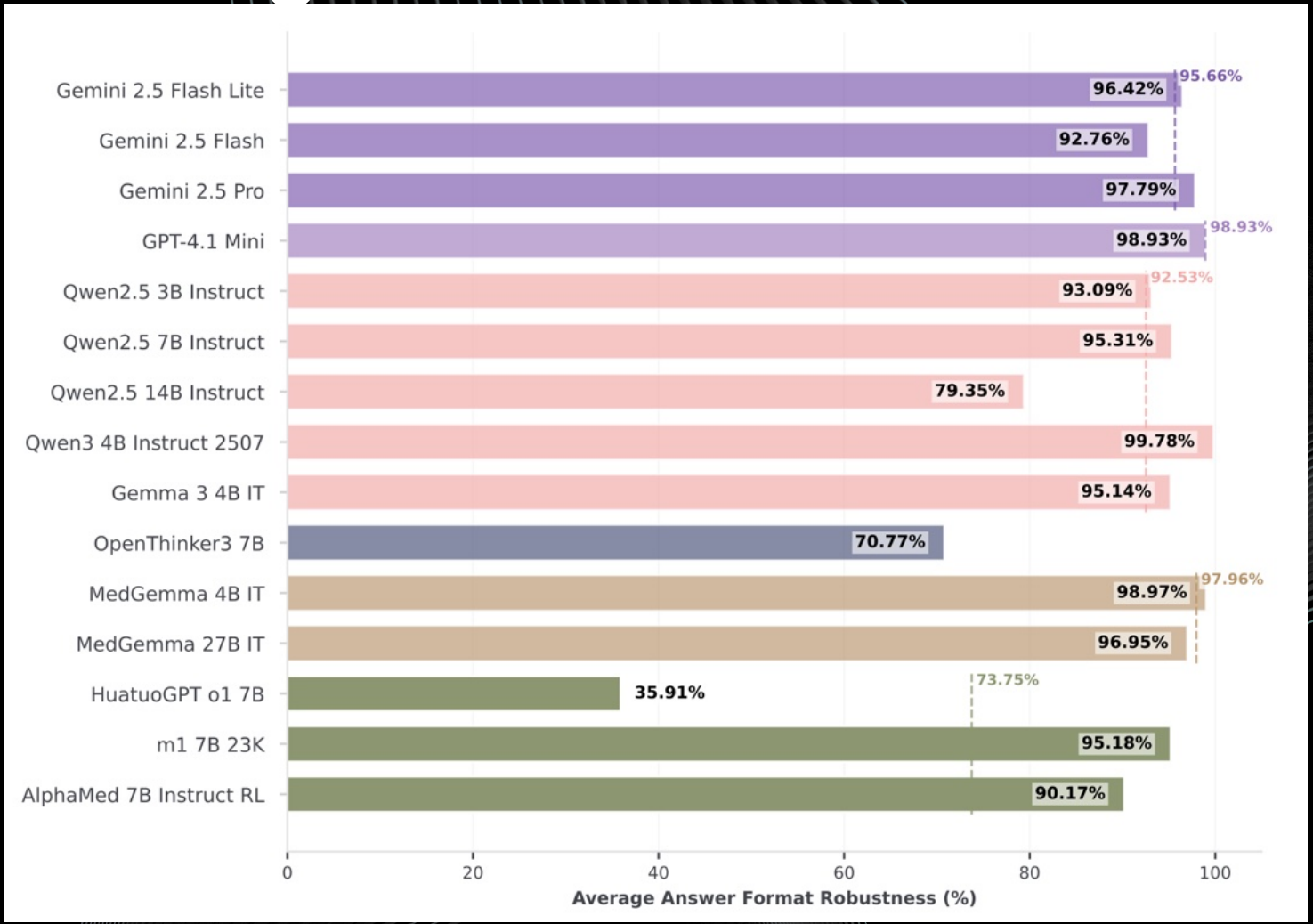
Robustness

Accuracy

Metrics

SiData+
Conference 2026

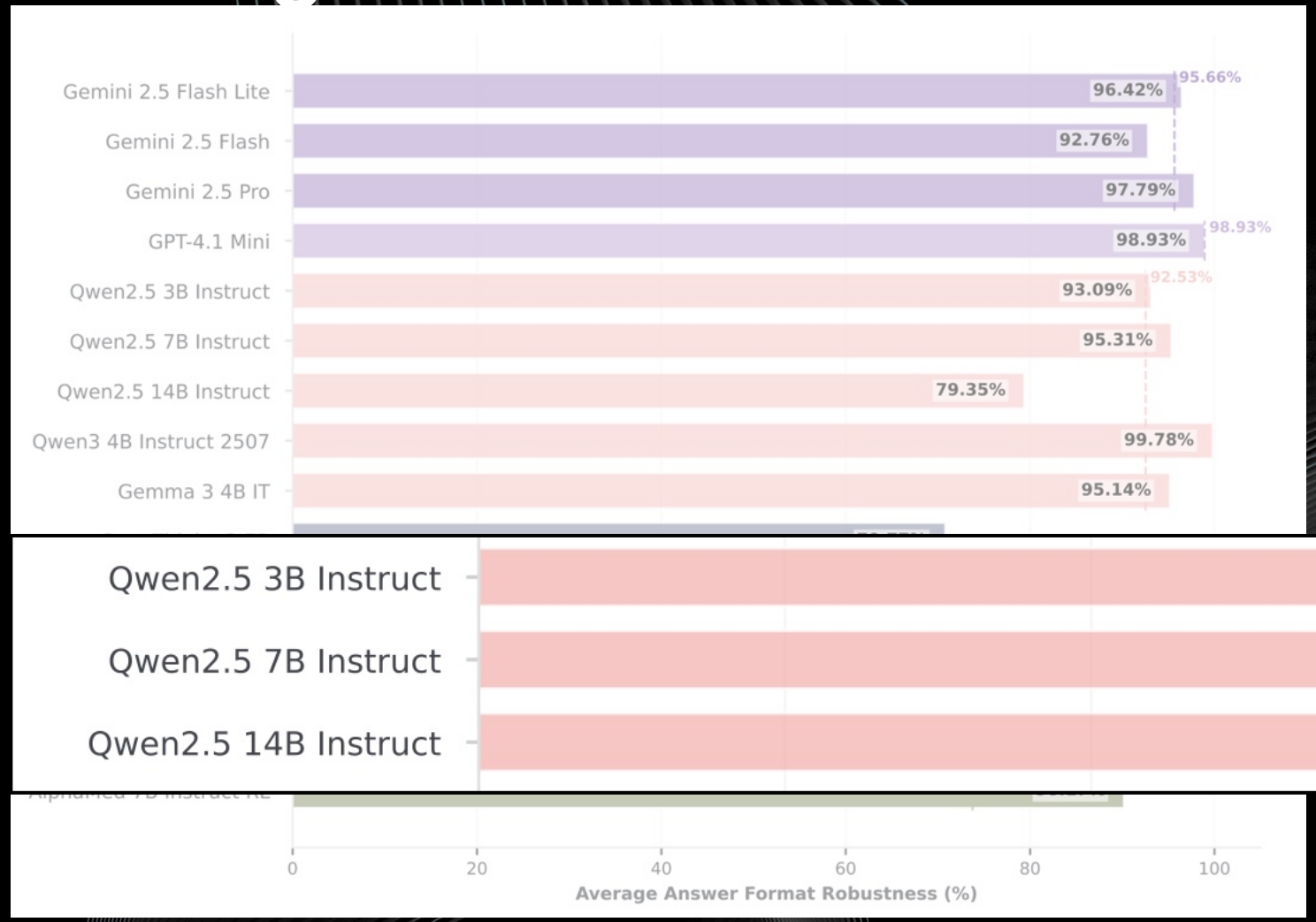
Findings



Finding 1
Most models are robust across different answer formats

SiData+
Conference 2026

Findings

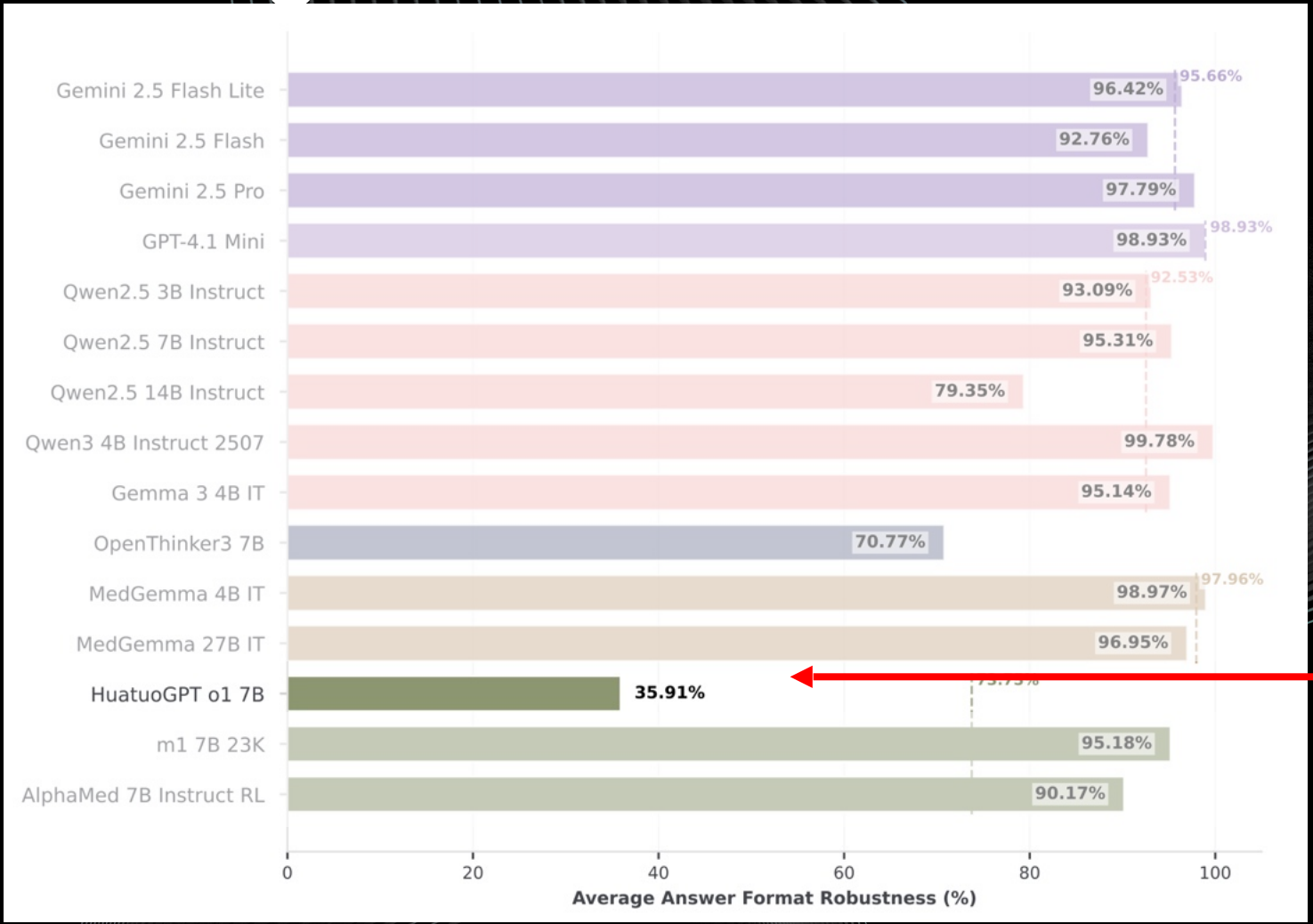


Finding 1
Most models are robust across different answer formats

Finding 2
Larger models are not necessarily better than smaller models

SiData+
Conference 2026

Findings



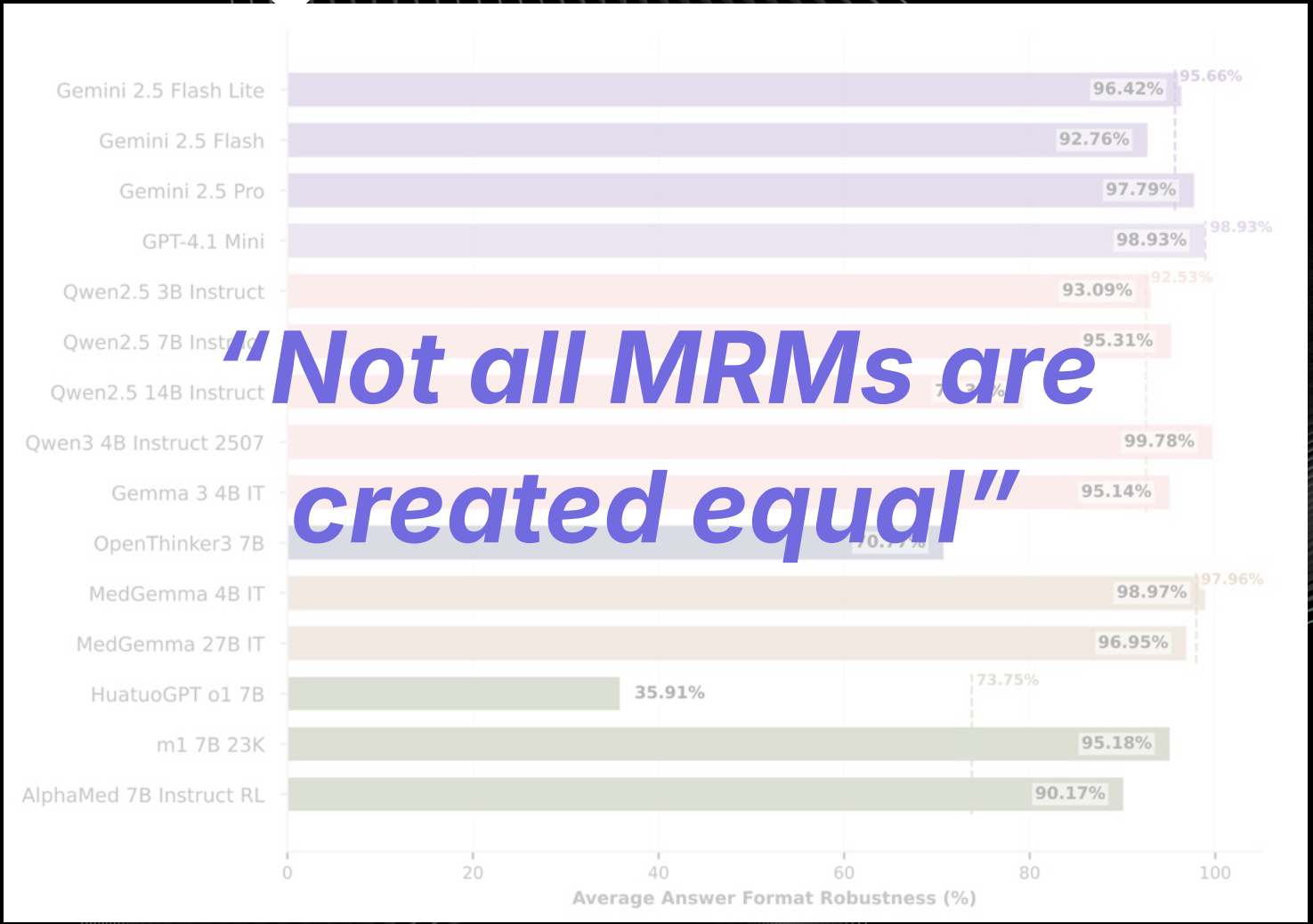
Finding 1
Most models are robust across different answer formats

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



"Not all MRMs are created equal"

Finding 1
Most models are robust across different answer formats

Finding 2
Larger models are not necessarily better than smaller models

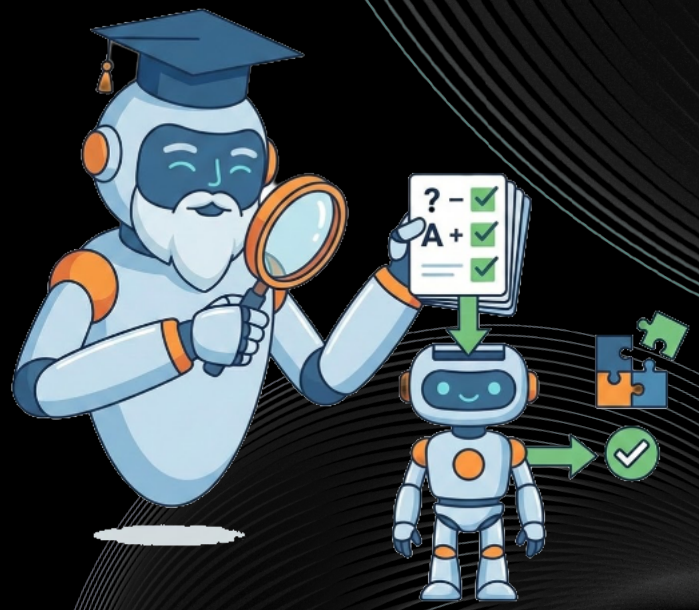
Finding 3
HuatuoGPT has the worst robustness, unlike other MRMs

SiData+
Conference 2026

How to Make a Model Think Longer?

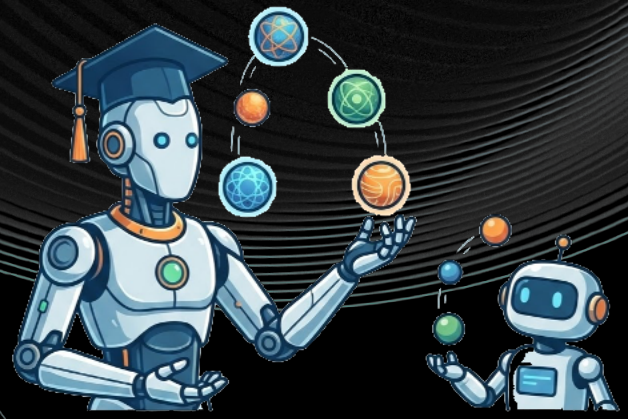
Teach It

Supervised Fine-Tuning



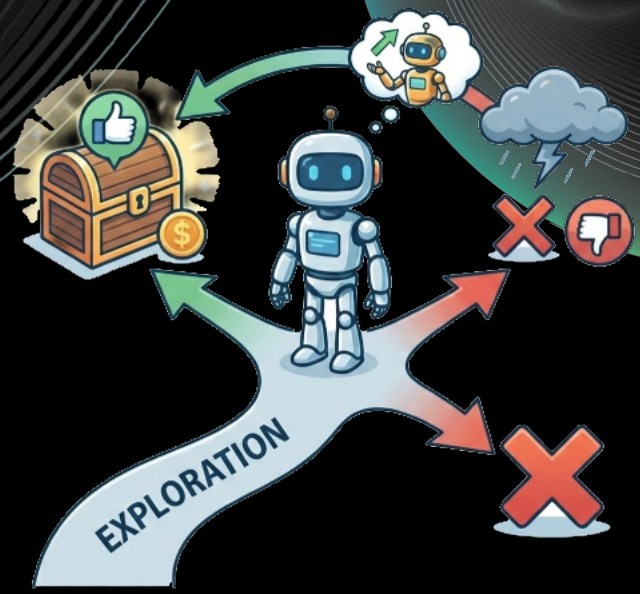
Ask It to Mimic

Knowledge Distillation



Let It Explore

Reinforcement Learning

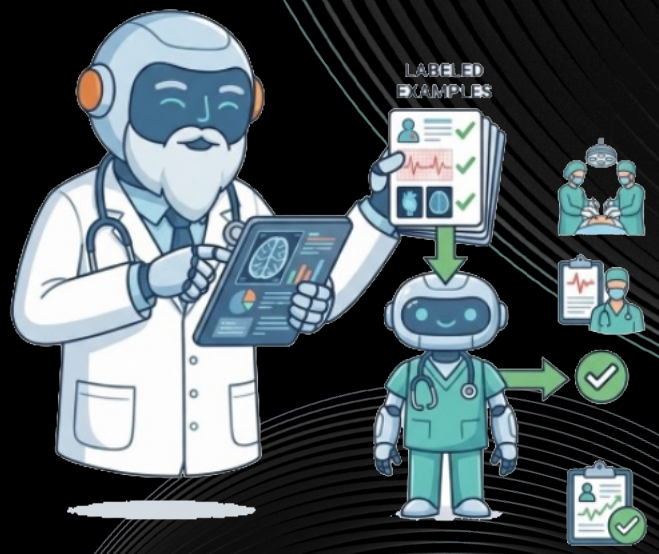


SiData+
Conference 2026

How to Make a Medical Reasoning Model?

Teach It

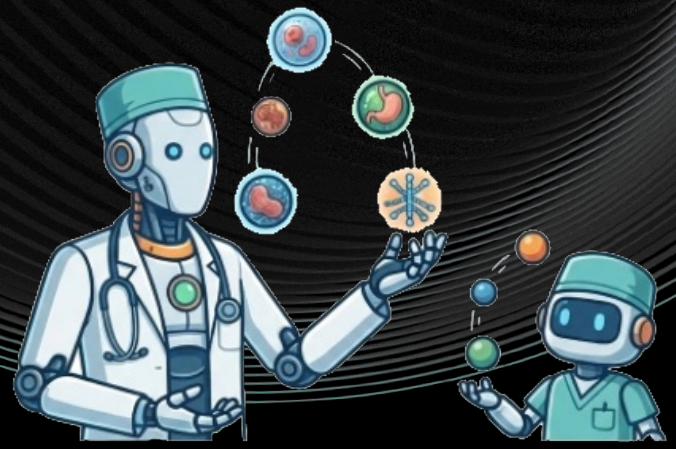
Teaching A Model Thinking Patterns



E.g., HuatuoGPT-o1

Ask It to Mimic

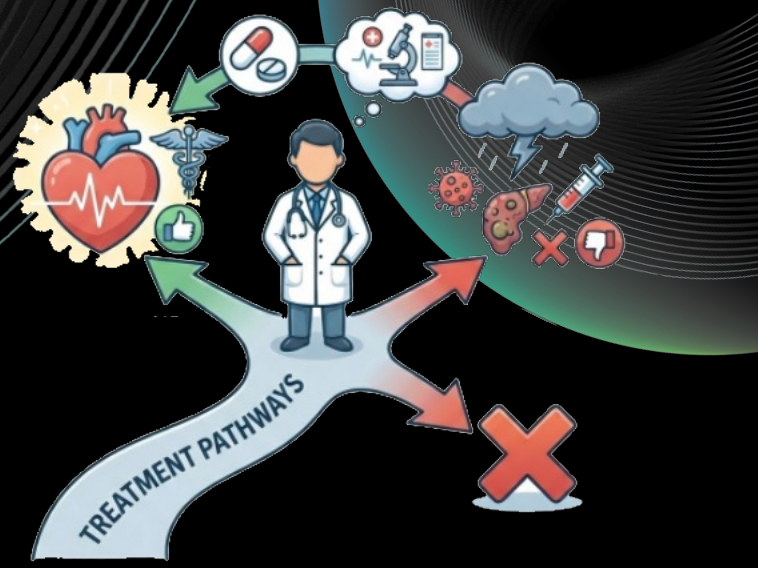
Imitating Teacher Thinking Patterns



E.g., m1

Let It Explore

Explore On Its Own How To Think



E.g., AlphaMed

Why Do We Need Controlled Experiments?

Problem

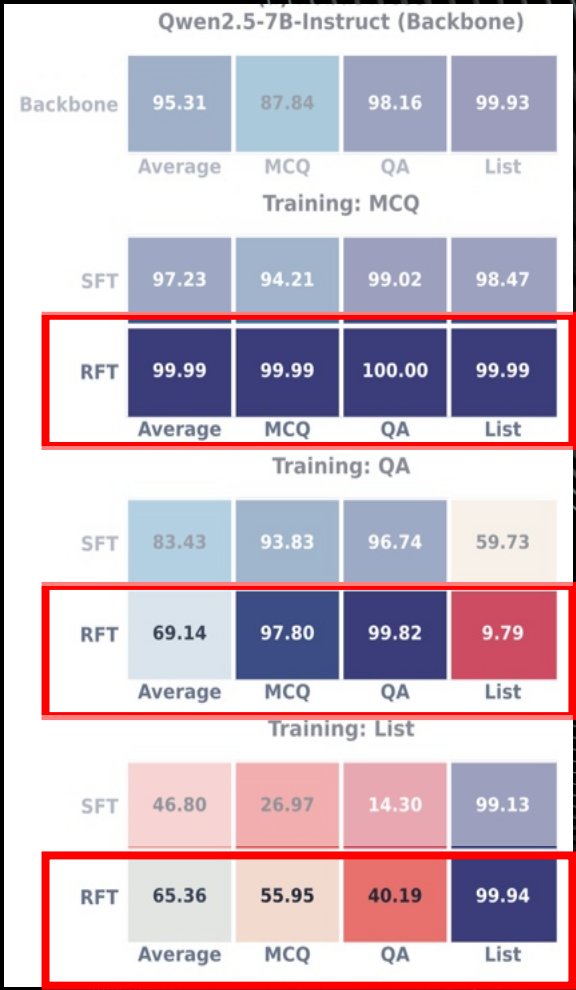
Comparing closed MRMs makes it very difficult to **identify training factors that affect robustness**

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)*

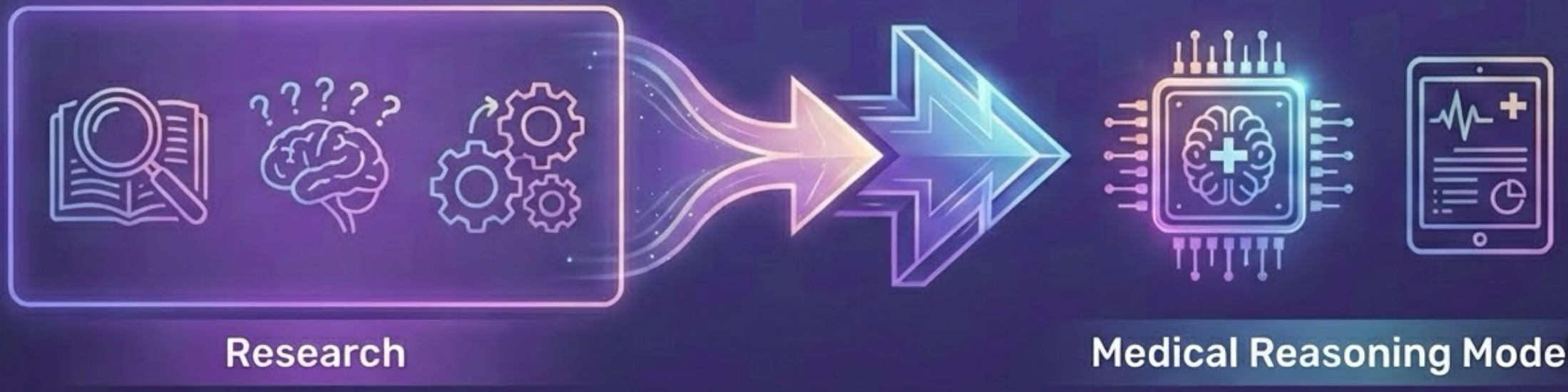
Training Paradigms (SFT vs. RFT) *(Finding 4)*

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 ~100× lower cost

Gemini 2.5 Pro is ~$11.25/1M tokens, while 4B model hosted on optimized inference is ~$0.10/1M tokens

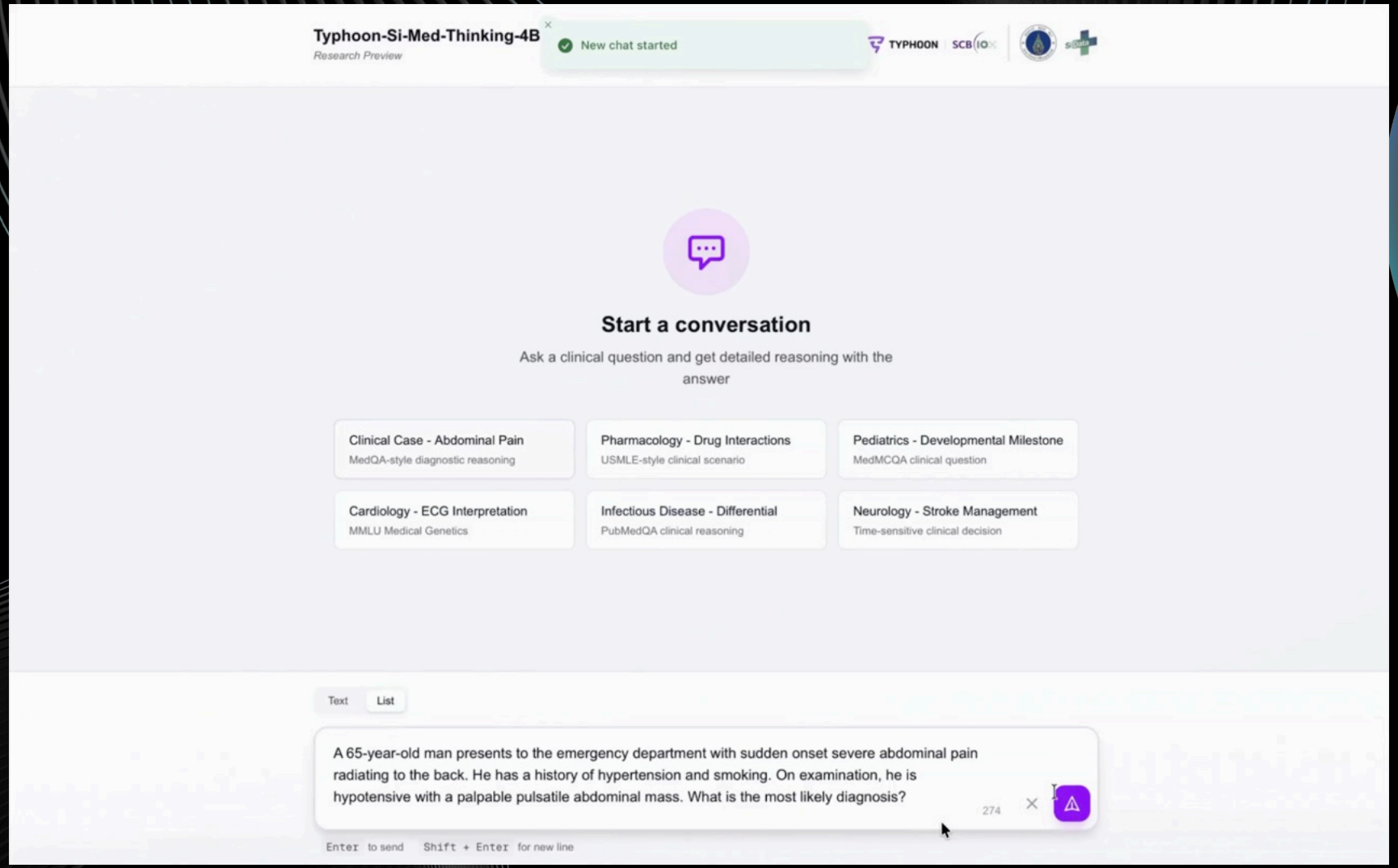
SiData+
Conference 2026

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	55.19	46.10	62.66	53.52
Gemini 2.5 Pro	**58.68**	**49.20**	68.46	58.85
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	48.87	**94.73**	**90.68**

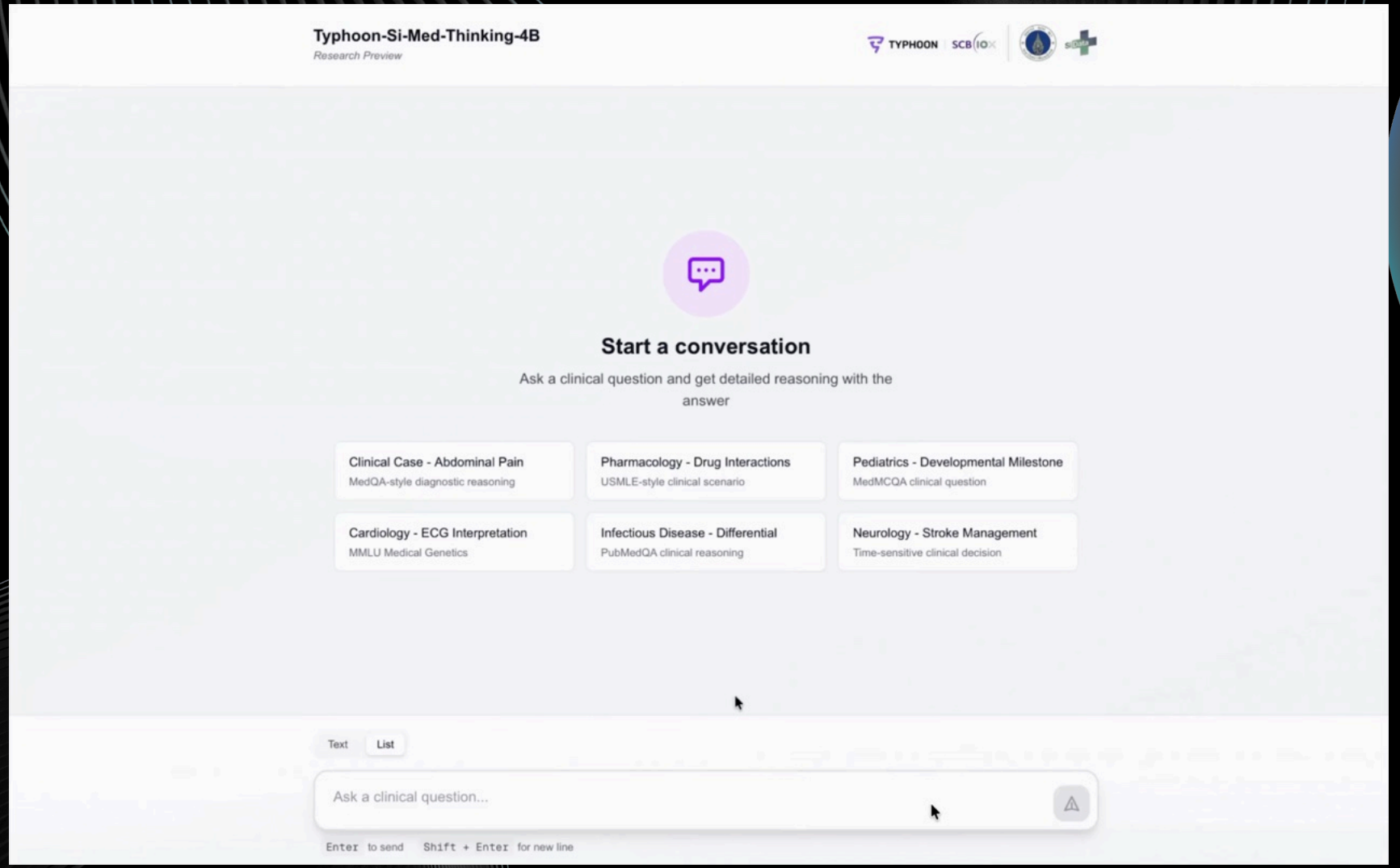
SiData+
Conference 2026

Demo: Ranked List Answer #1



SiData+
Conference 2026

Demo: Ranked List Answer #2



SiData+
Conference 2026

Open Knowledge & Open Model

📚 **Preprint Paper**

Models Are Available
On 🤗 **Hugging Face**

SiData+
Conference 2026

SiData+
Conference 2026

Supervised Fine-Tuning with Synthetic Data

Prompts

Distilled Reasoning Model

Sampling

SFT

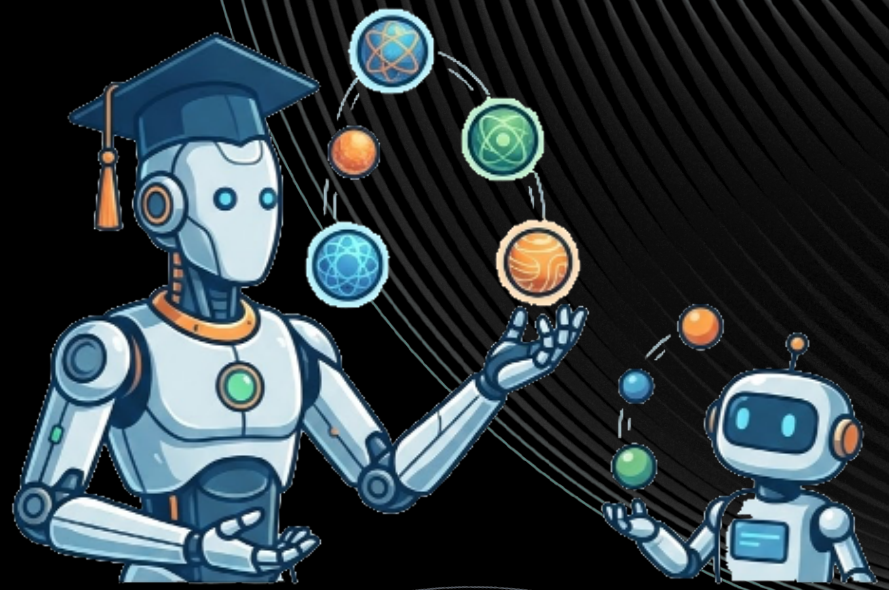
Student Model

Synthetic Data Agentic WorkFlow

Generated **Long Thought** Responses

SiData+
Conference 2026

Knowledge Distillation



Prompts

Sampling

Reasoning Model

Generated **Long Thought** Responses

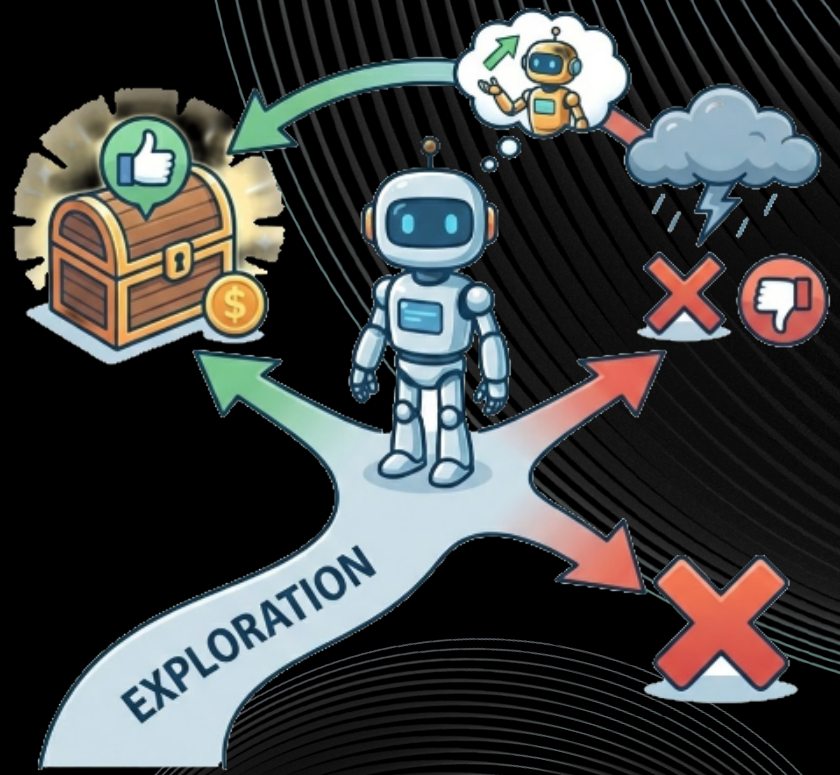
SFT

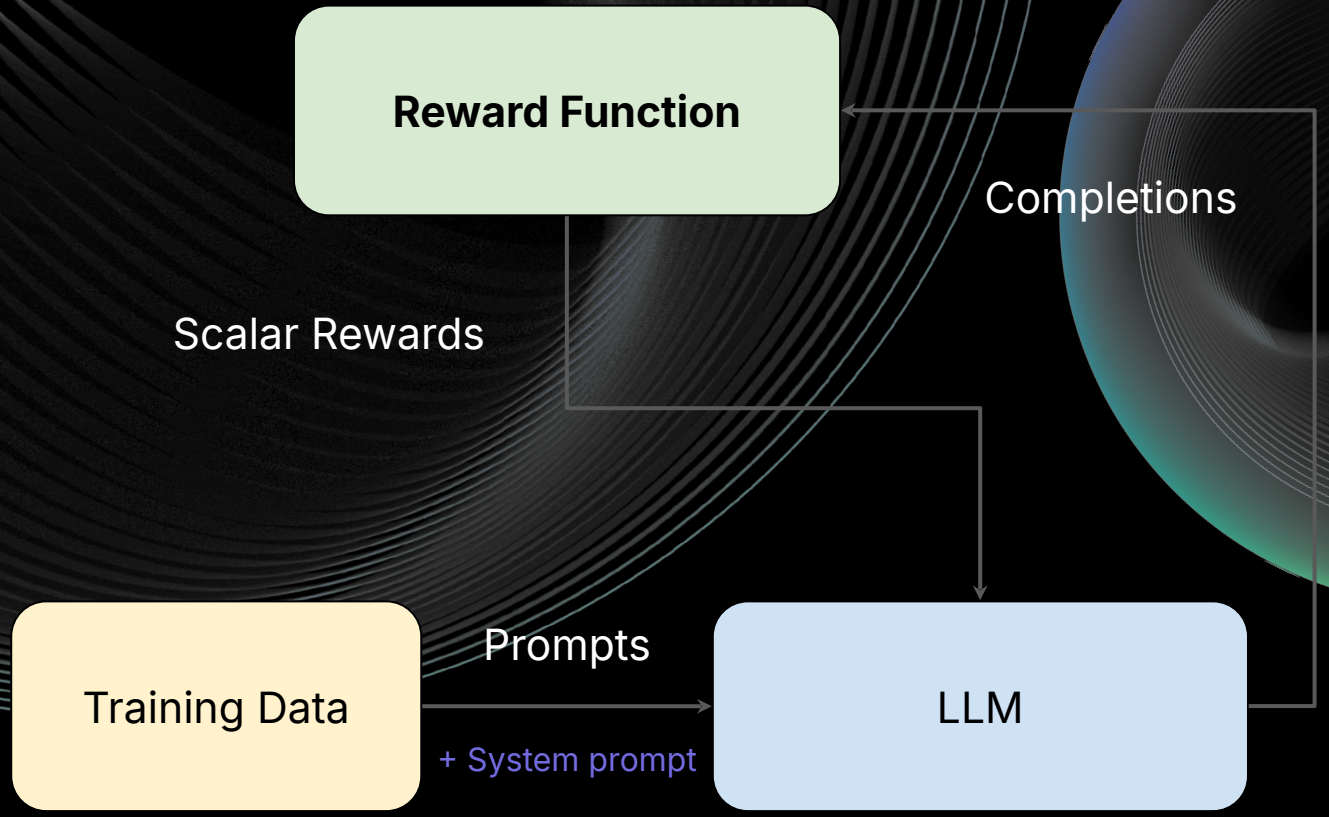
Student Model

Distilled Reasoning Model

SiData+
Conference 2026

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>. [...]*