From Illusion to Robust Data

Bae Sun Woo

December 29, 2025

“Do LRMs really reason or do they just pretend to do so?

Table of Contents

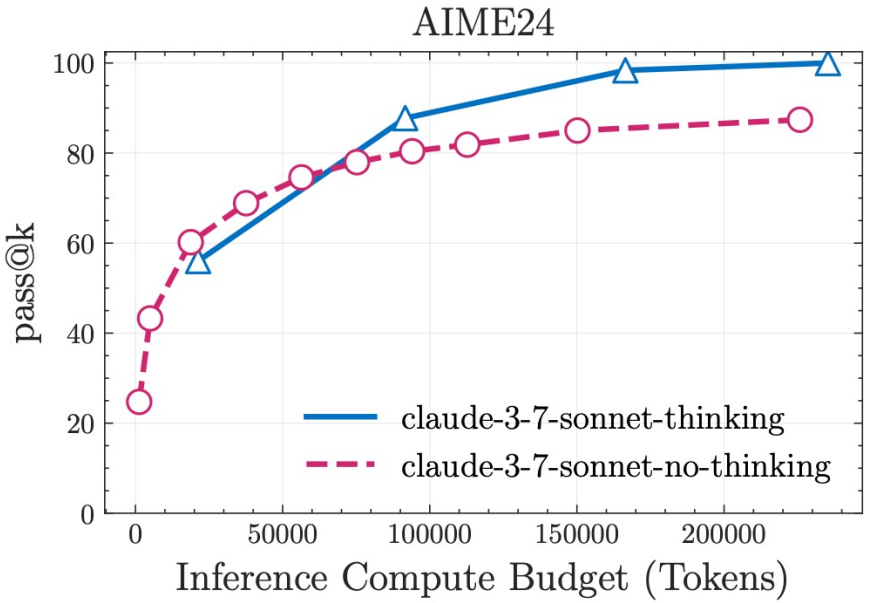
The Illusion of Thinking - Motivation

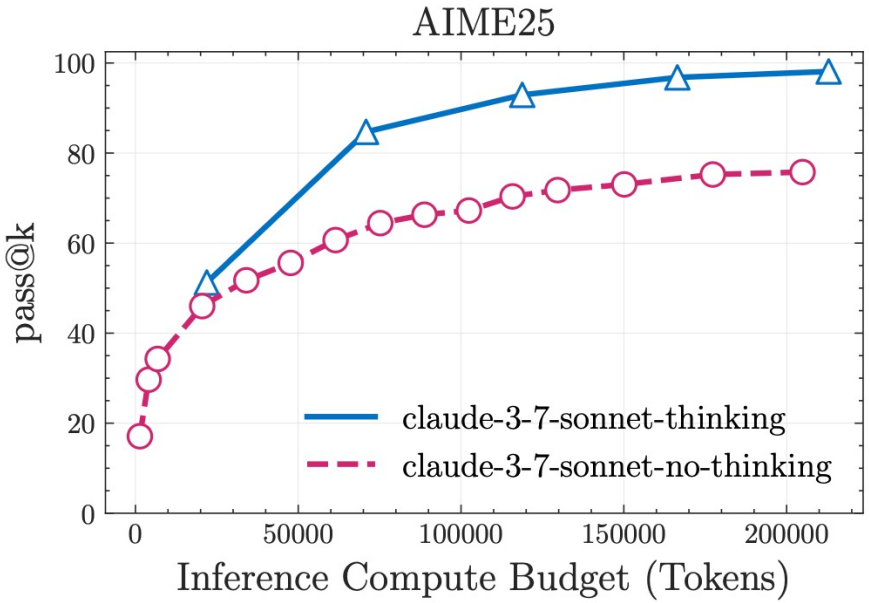
“Are these models capable of generalizable reasoning, or are they leveraging different forms of pattern matching?”

“How does their performance scale with increasing problem complexity?”

“What are the inherent limitations of current reasoning approaches, and what improvements might be necessary to advance toward more robust reasoning capabilities?”

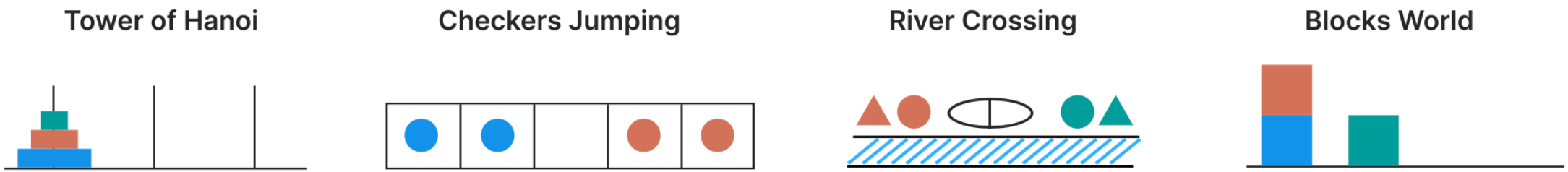
Current Problems in Benchmark





- Established Math Benchmark: MATH500, AIME24, AIME25
- Data contamination : Human(AIME24) < Human(AIME25), LM(AIME24) > LM(AIME25)
- Cannot control problem complexity
- pass@k: probability that at least one correct solution is found among k independent attempts

Instead, Controllable Puzzle Environments



1. Controllable Complexity

Puzzle	Tower of Hanoi	Checkers Jumping	River Crossing	Blocks World
Controllable N	Disks	Checkers	Actor/Agent Pairs	Blocks

2. Low Data Contamination

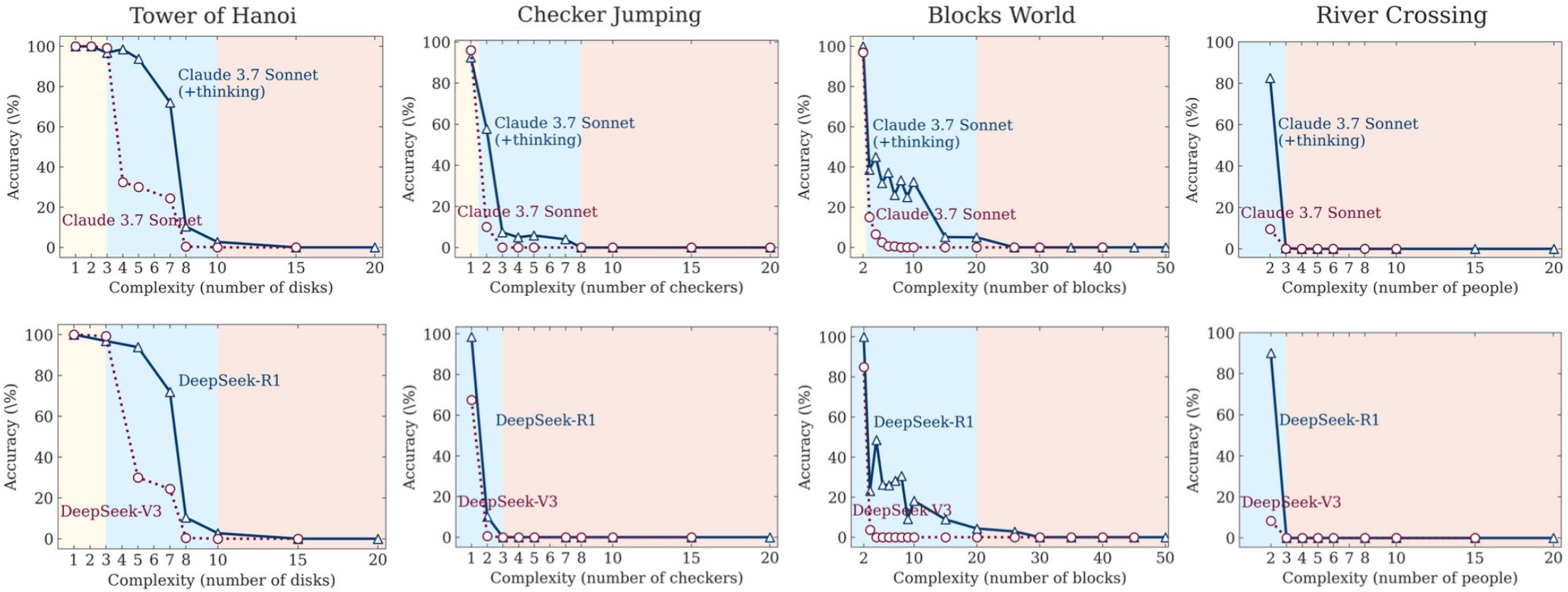
Experiment Setup

Thinking Model	Non – Thinking Counterpart
Claude 3.7 Sonnet (+ thinking)	Claude 3.7 Sonnet
DeepSeek-R1	DeepSeek-V3

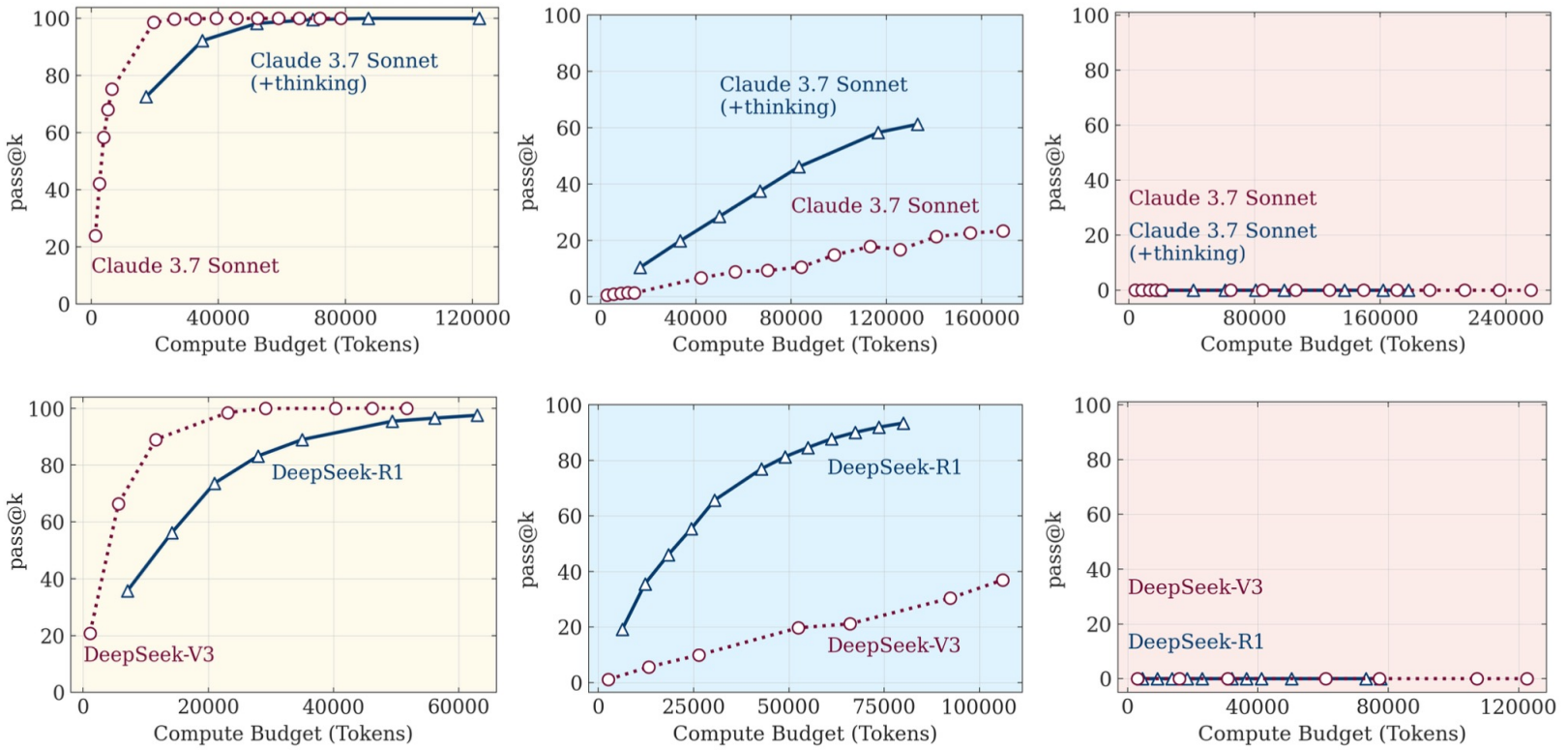
* Above models are chosen due to their accessibiliy to thinking tokens

For each puzzle instance, generate 25 samples and report the average performance of each model across them.

Result 1 : Three Regimes of Complexity - Accuracy



Result 1 : Three Regimes of Complexity – Token Efficiency

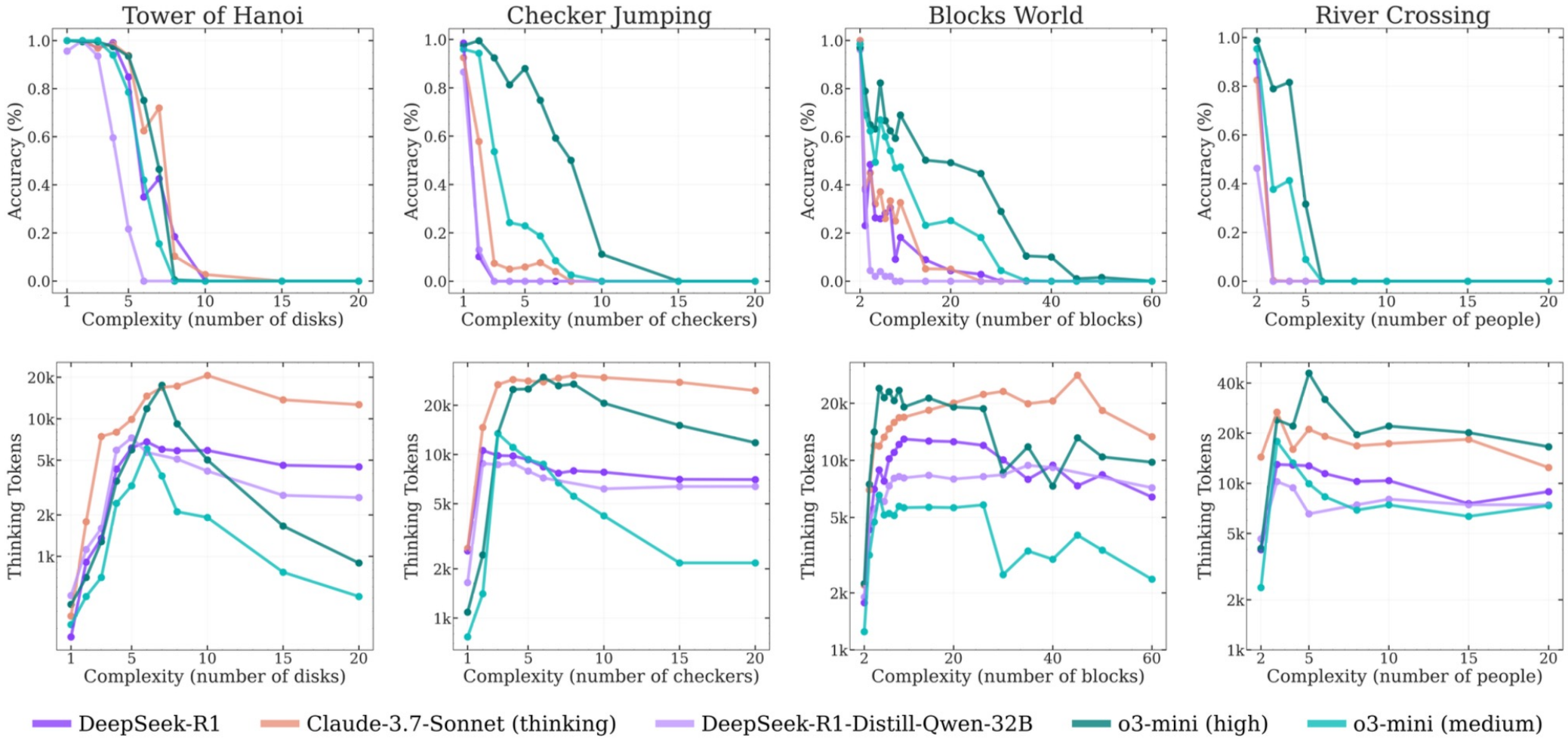


Yellow : Low Complexity Blue : Medium Complexity Red : High Complexity

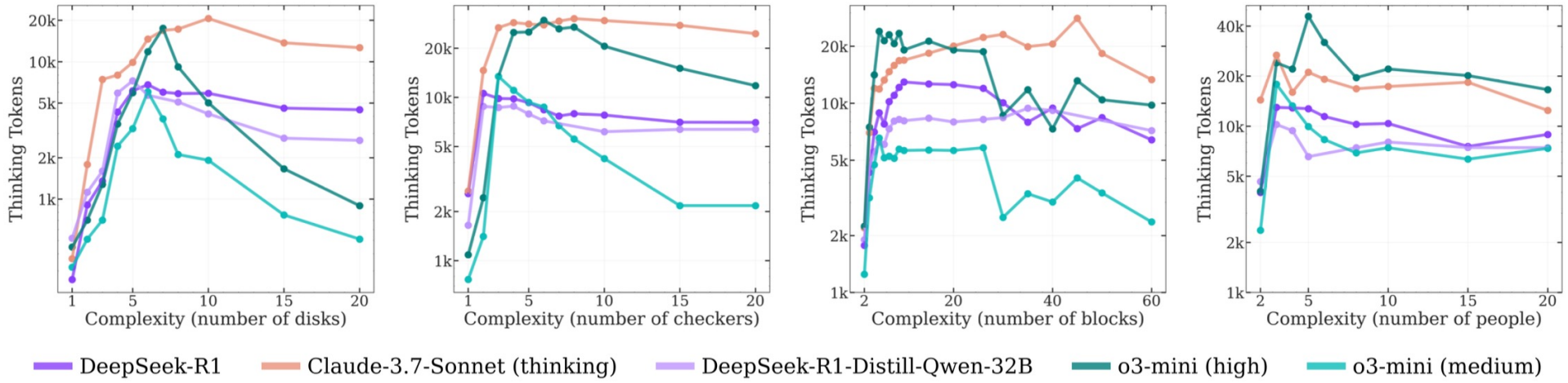
Result 1 : Three Regimes of Complexity

Complexity		Low	Medium	High
Thinking Models	**Accuracy**	Comparable	Better Performance (∵ long CoT)	Collapse to Zero
	Token Efficiency	Less Token Efficient	-	
Non Thinking Models	**Accuracy**	Comparable	Worse performance	Collapse to Zero
	Token Efficiency	More Token Efficient	-	

Result 2 : Collapse of Reasoning Model



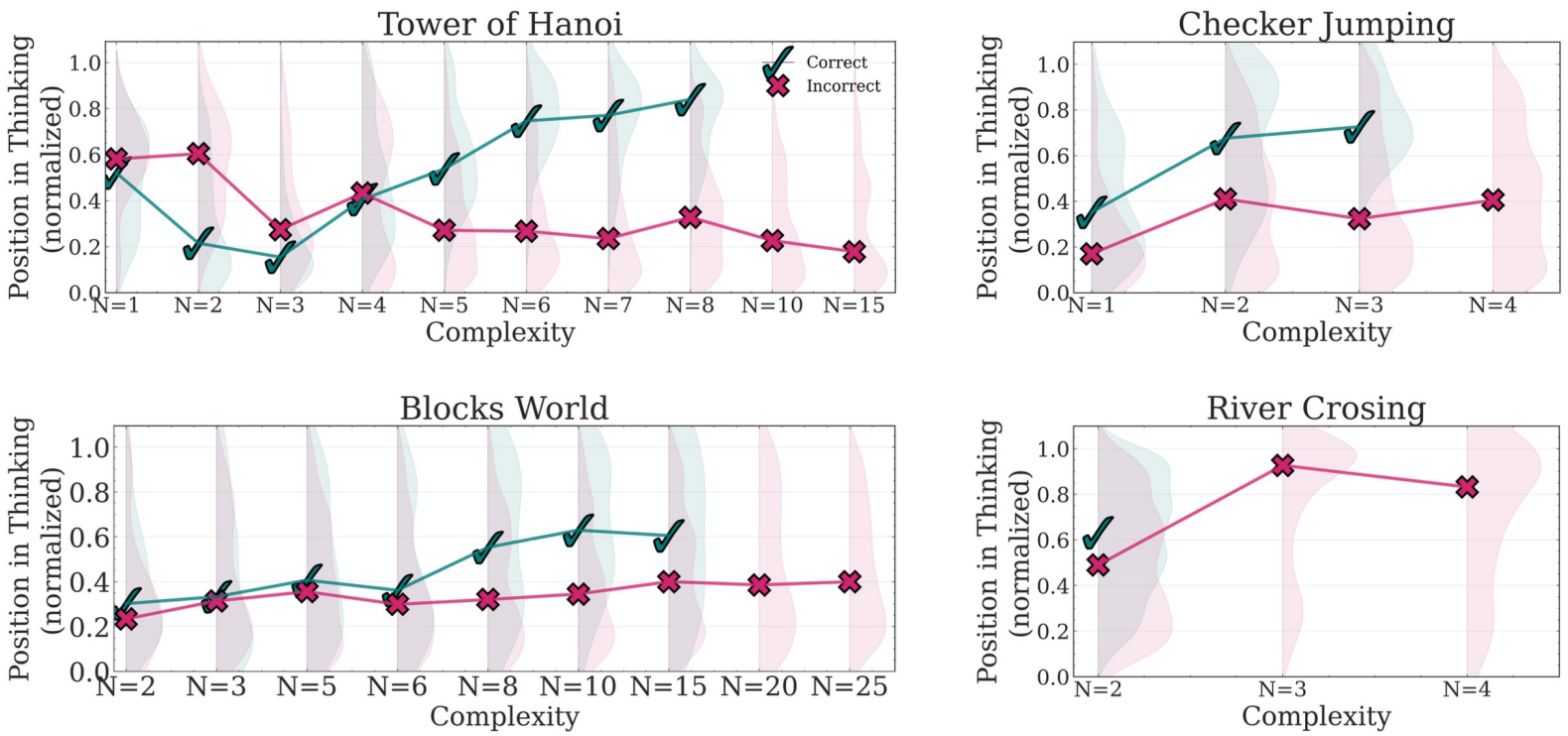
Result 2 : Collapse of Reasoning Model - Token



Approaching a critical threshold, models counterintuitively begin to reduce their reasoning effort (thinking tokens) despite increasing problem difficulty.

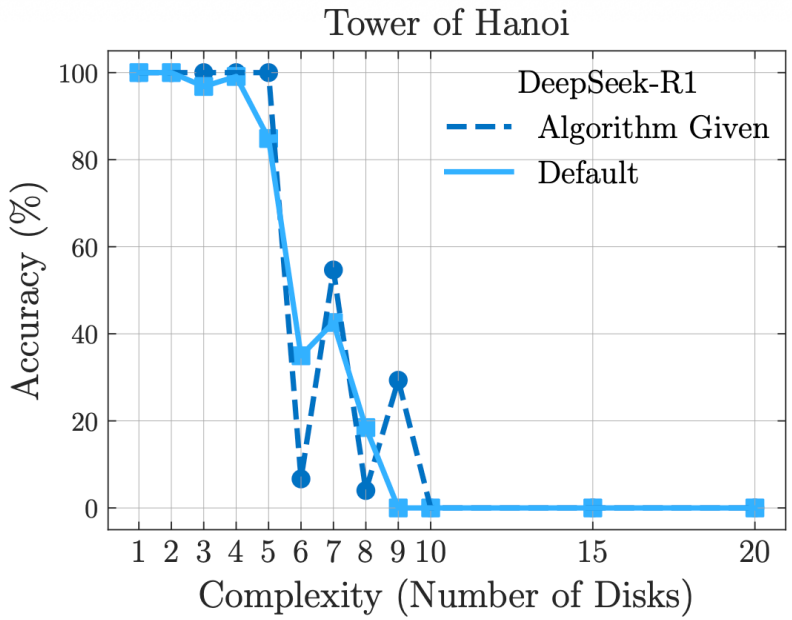
“Fundamental scaling limitation in the thinking capabilities”

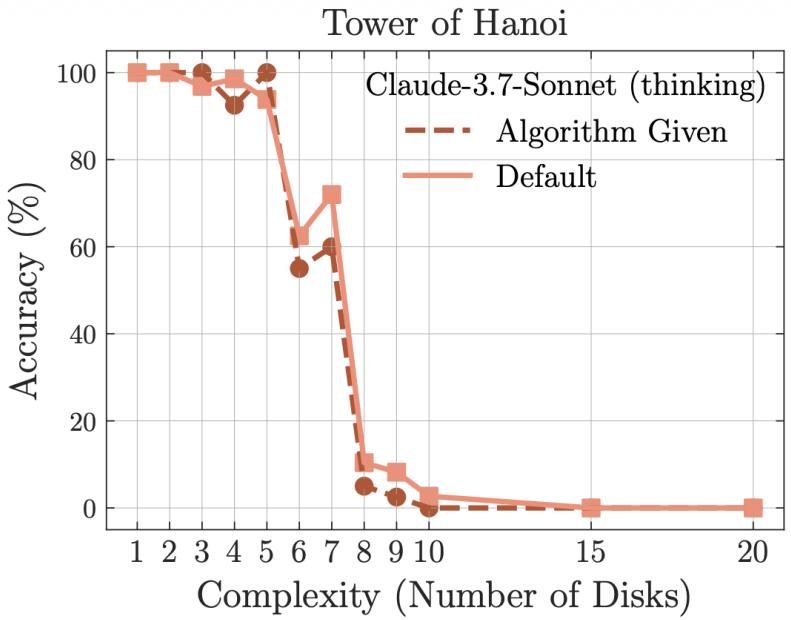
Result 3 : Overthinking in Low Complexity



Overthinking => Waste of Compute

Result 4 : Puzzling Behavior of Reasoning Models





Algorithm is given => Still Collapse => Limitations in verification and in following logical steps

The Illusion of Thinking - Conclusion

- Collapse in High Complexity
- Fundamental Scaling Limitation in the Thinking Capabilities
- Overthinking
- Limitations in Verification and in following Logical Steps

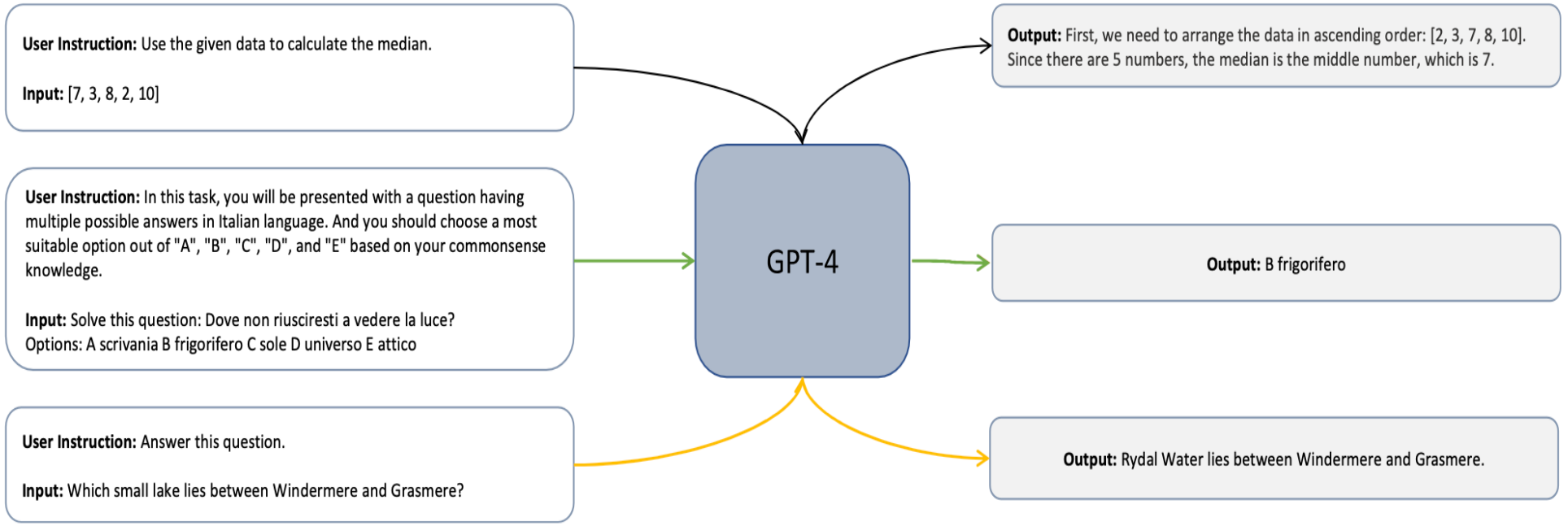
“Lack of Generalizable Reasoning Ability”

“Can reasoning ability be improved through better data?”

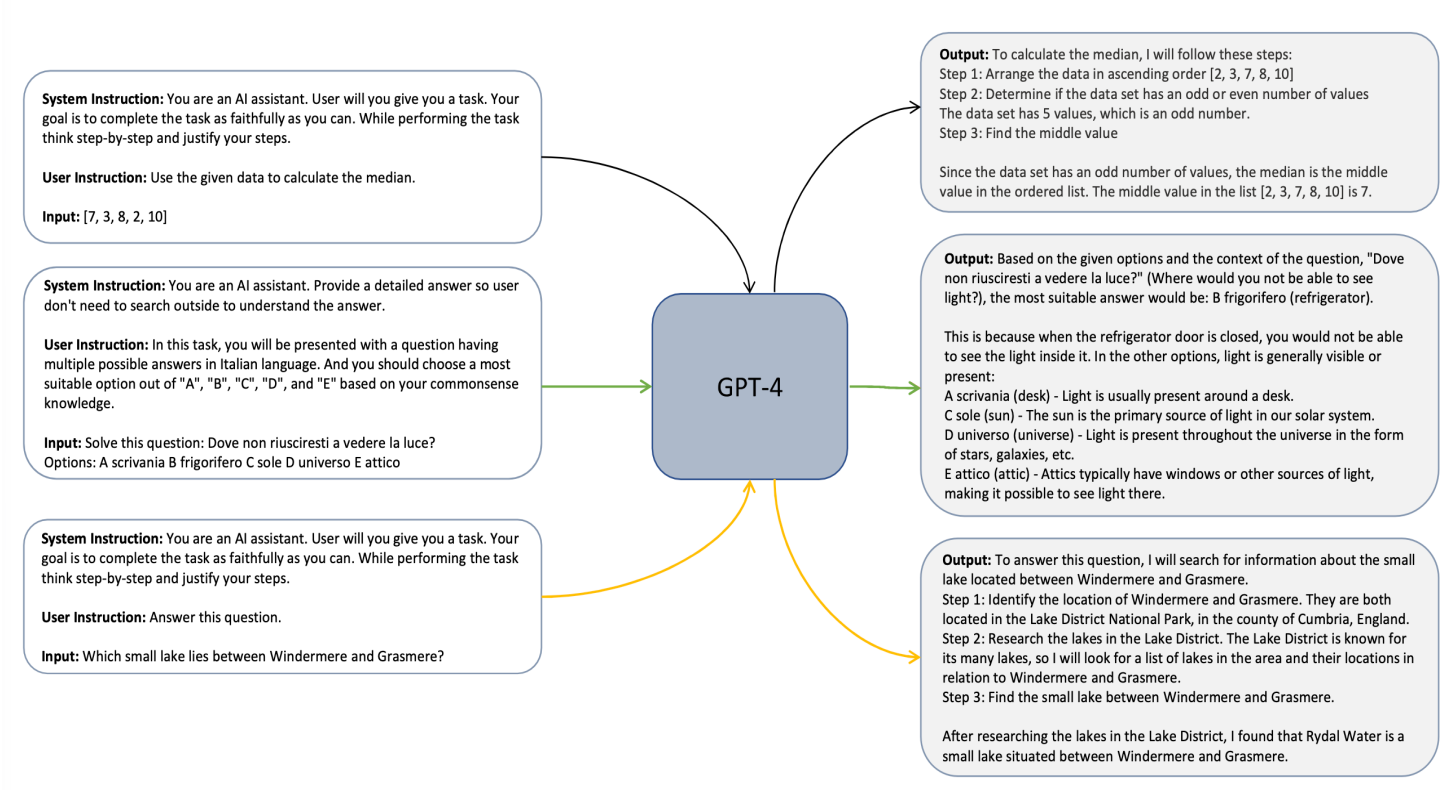
ORCA – Motivation (Current Problems)

	Exsiting Small Model (Vicuna - 13B)	Orca - 13B
Instruction	Simple Instruction Tuning <query, response> ⇒ Lack of Reasoning Traces	Explanation Tuning (System Instruction) <System message, query, response > => Reasoning Traces
Task Diversity	ShareGPT	Flan 2022 Collection (extensive public assortment of tasks and instructions)
Evaluation	GPT-4 auto-evaluation : Biased	GPT-4 auto-evaluation + Academic Benchmark + Professional/Academic Exams + Safety evaluation

Existing Instruction Tuning



Core Idea : Explanation Tuning



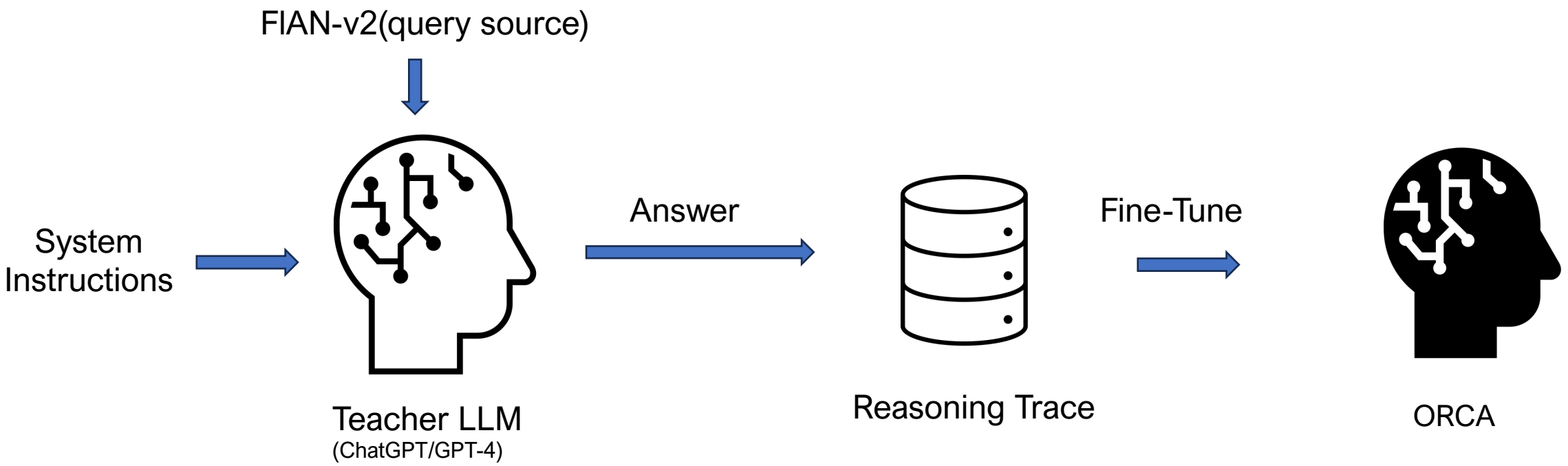
Add a system instruction to shape the teacher into a “good explainer”.

=> Output Includes Reasoning Traces

System Instructions

Id.	System Message
1	<empty system message>
2	You are an AI assistant. Provide a detailed answer so user don't need to search outside to understand the answer.
3	You are an AI assistant. You will be given a task. You must generate a detailed and long answer.
4	You are a helpful assistant, who always provide explanation. Think like you are answering to a five year old.
5	You are an AI assistant that follows instruction extremely well. Help as much as you can.
6	You are an AI assistant that helps people find information. Provide a detailed answer so user don't need to search outside to understand the answer.
7	You are an AI assistant. User will you give you a task. Your goal is to complete the task as faithfully as you can. While performing the task think step-by-step and justify your steps.
8	You should describe the task and explain your answer. While answering a multiple choice question, first output the correct answer(s). Then explain why other answers are wrong. Think like you are answering to a five year old.
9	Explain how you used the definition to come up with the answer.
10	You are an AI assistant. You should describe the task and explain your answer. While answering a multiple choice question, first output the correct answer(s). Then explain why other answers are wrong. You might need to use additional knowledge to answer the question.
11	You are an AI assistant that helps people find information. User will you give you a question. Your task is to answer as faithfully as you can. While answering think step-by-step and justify your answer.
12	User will you give you a task with some instruction. Your job is follow the instructions as faithfully as you can. While answering think step-by-step and justify your answer.
13	You are a teacher. Given a task, you explain in simple steps what the task is asking, any guidelines it provides and how to use those guidelines to find the answer.
14	You are an AI assistant, who knows every language and how to translate one language to another. Given a task, you explain in simple steps what the task is asking, any guidelines that it provides. You solve the task and show how you used the guidelines to solve the task.
15	Given a definition of a task and a sample input, break the definition into small parts. Each of those parts will have some instruction. Explain their meaning by showing an example that meets the criteria in the instruction. Use the following format: Part #: a key part of the definition. Usage: Sample response that meets the criteria from the key part. Explain why you think it meets the criteria.
16	You are an AI assistant that helps people find information.

Orca Overview



Training Setup & Evaluation Protocol

- Model: 13B pretrained backbone(LLaMA) → supervised fine-tuning (SFT) on Orca's explanation dataset.
- Evaluation focuses on reasoning-heavy and academic/professional benchmarks:
  - BigBench Hard (BBH) – complex zero-shot reasoning
  - AGIEval – exam-style tasks

Key Result

Task	Human -Avg	Human -Top	TD-003	Chat GPT	GPT-4	Vicuna-13B	Orca-13B
AQuA-RAT	85	100	29.9	31.9	40.6	20.1	**27.9** (39.2%)
LogiQA	86	95	22.7	35	49.3	29.8	**35.2** (18.1%)
LSAT-AR	56	91	21.7	24.4	35.2	20.4	**21.3** (4.3%)
LSAT-LR	56	91	47.5	52.6	80.6	32.6	**43.9** (34.9%)
LSAT-RC	56	91	64.7	65.4	85.9	32.7	**57.3** (75.0%)
SAT-Math	66	94	35.5	42.7	64.6	28.6	**32.3** (12.7%)
SAT-English	66	94	74.8	81.1	88.8	44.2	**76.7** (73.6%)
SAT-English (w/o Psg.)	66	94	38.4	44.2	51	26.2	**38.8** (48.1%)
Average	67.1	93.8	41.9	47.2	62	29.3	**41.7** (42.1%)

AGIEval benchmark

Task	ChatGPT	GPT-4	Vicuna-13B	Orca-13B
Boolean Expressions	82.8	77.6	40.8	**72.0** (76.5%)
Causal Judgement	57.2	59.9	42.2	**59.9** (41.8%)
Date Understanding	42.8	74.8	10.0	**50.0** (400.0%)
Disambiguation QA	57.2	69.2	18.4	**63.6** (245.7%)
Formal Fallacies	53.6	64.4	47.2	**56.0** (18.6%)
Geometric Shapes	25.6	40.8	3.6	**20.8** (477.8%)
Hyperbaton	69.2	62.8	44.0	**64.0** (45.5%)
Logical Deduction (5 objects)	38.8	66.8	4.8	**39.6** (725.0%)
Logical Deduction (7 objects)	39.6	66.0	1.2	**36.0** (2900.0%)
Logical Deduction (3 objects)	60.4	94.0	16.8	**57.6** (242.9%)
Movie Recommendation	55.4	79.5	43.4	**78.3** (80.6%)
Navigate	55.6	68.8	46.4	**57.6** (24.1%)
Penguins in a Table	45.9	76.7	15.1	**42.5** (181.8%)
Reasoning about Colored Objects	47.6	84.8	12.0	**48.4** (303.3%)
Ruin Names	56.0	89.1	15.7	**39.5** (151.2%)
Salient Translation Error Detection	40.8	62.4	2.0	**40.8** (1940.0%)
Snarks	59.0	87.6	28.1	**62.4** (122.0%)
Sports Understanding	79.6	84.4	48.4	**67.2** (38.8%)
Temporal Sequences	35.6	98.0	16.0	**72.0** (350.0%)
Tracking Shuffled Objects (5 objects)	18.4	25.2	9.2	**15.6** (69.6%)
Tracking Shuffled Objects (7 objects)	15.2	25.2	5.6	**14.0** (150.0%)
Tracking Shuffled Objects (3 objects)	31.6	42.4	23.2	**34.8** (50.0%)
Web of Lies	56.0	49.6	41.2	**51.2** (24.3%)
Average	48.9	67.4	23.3	**49.7** (113.7%)

BBH benchmark

Main Limitations

Data Bias: Systematic bias inherited and amplified from teacher-generated explanation traces

Hallucination: Teacher-generated reasoning traces may contain unverified or logically incorrect steps

“What if a lightweight agent could actively verify, critique, and probe the teacher’s reasoning?”