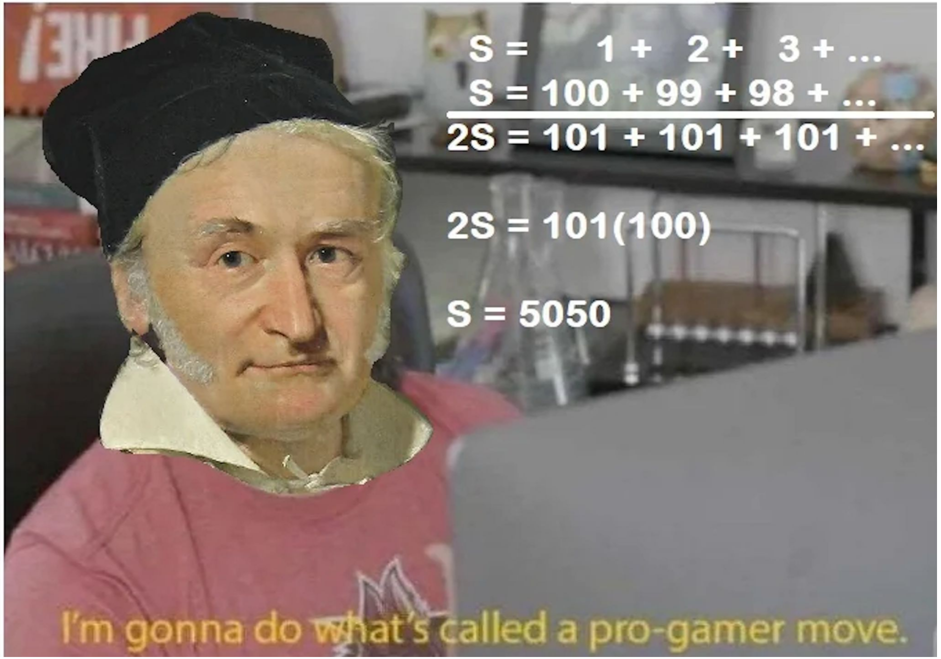
Creativity or Brute Force? Using Brainteasers as a Window into the Problem-Solving Abilities of Large Language Models

Beyond final answer accuracy, we holistically evaluate multiple layers of the reasoning process

teacher:*tells students to sum natural numbers from 1 to 100 thinking it'll take them a long time*
Gauss:


S = 1 + 2 + 3 + ...
S = 100 + 99 + 98 + ...
2S = 101 + 101 + 101 + ...
2S = 101(100)
S = 5050
I'm gonna do what's called a pro-gamer move.

01 Creative Solution Definition

02 Braingle Brainteasers Benchmark

03 Frontier Reasoning Models' Performance

04 Frontier Reasoning Models' Creativity/Brute-force Rate

05 Other takeaways

Creative Solution Definition

We define a creative solution as an innovative, insight-driven approach that leverages pattern recognition or lateral thinking.

Rather than exhaustively testing all possibilities,
it reframes the problem or exploits shortcuts to reduce complexity.

Such solutions often involve minimal computation and are especially valuable for problems where brute-force search is intractable.

Braingle Brainteasers Benchmark

- **Content:** 478 problems (242 math, 236 logic) from Braingle, curated for style difficulty & reasoning diversity.

- **Features:** hints provided, human solution available, multiple solving strategies available (creative vs brute-force)

Braingle Brainteasers Benchmark

- **Focus:** Minimal knowledge barrier → tests reasoning rather than recall.

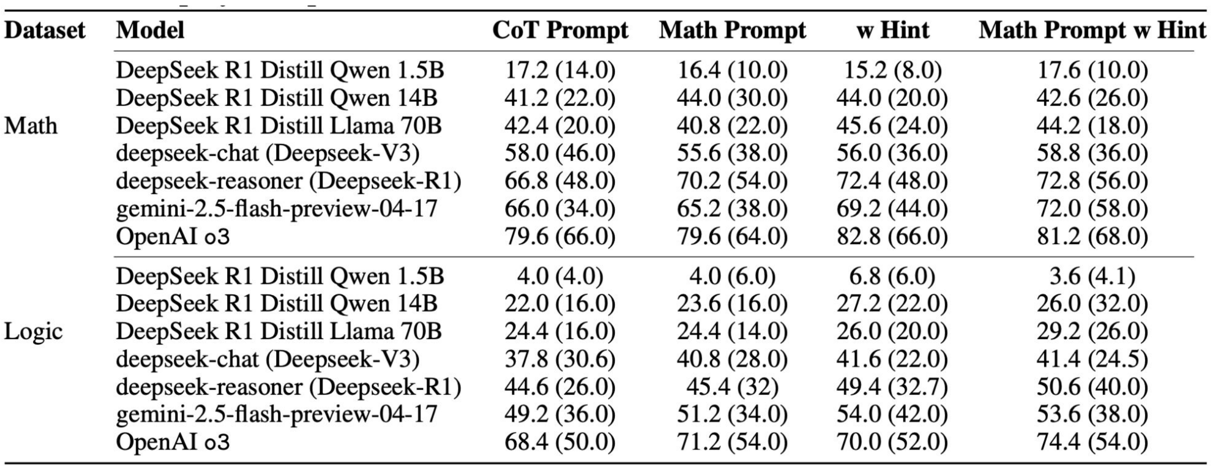
- **Complimentary to existing datasets**: providing a middle ground between structured logic datasets such as Zebra-Logic and formal mathematical ones such as MATH and AIME-2025.

How do LLMs perform on Brainteasers?

The **CoT Prompt** encourages the model to generate a **step-by-step** solution.

The **Math Prompt** additionally encourages the model to use **rigorous mathematical reasoning**, explicitly discouraging brute force, guesswork, and shortcuts.

How do LLMs perform on Brainteasers?

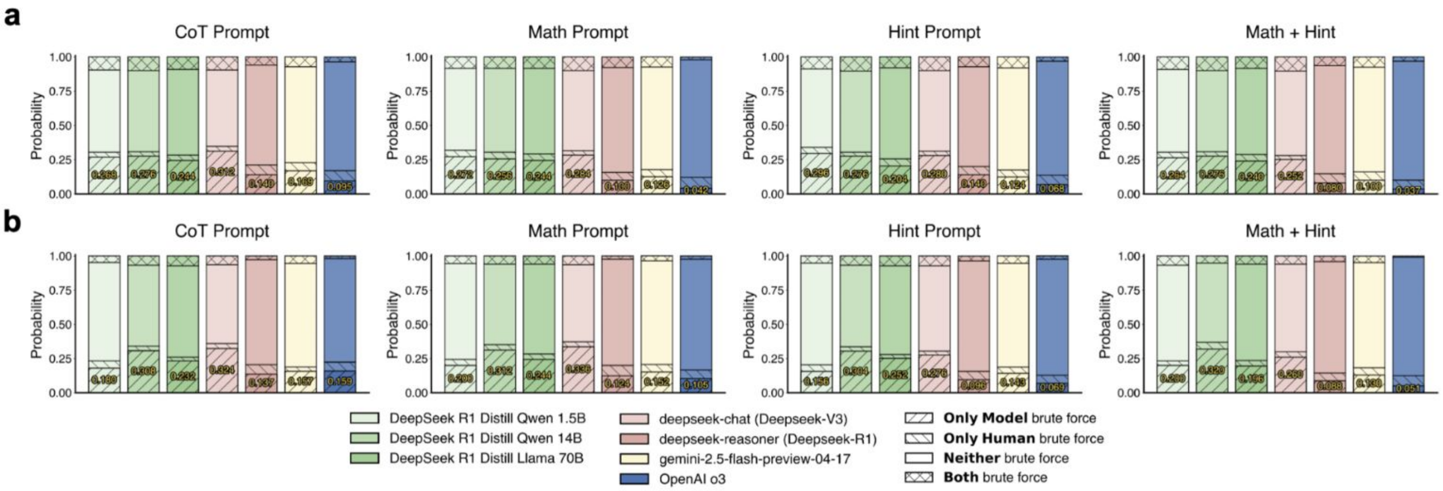


Dataset	Model	CoT Prompt	Math Prompt	w Hint	Math Prompt w Hint
Math	DeepSeek R1 Distill Qwen 1.5B	17.2 (14.0)	16.4 (10.0)	15.2 (8.0)	17.6 (10.0)
	DeepSeek R1 Distill Qwen 14B	41.2 (22.0)	44.0 (30.0)	44.0 (20.0)	42.6 (26.0)
	DeepSeek R1 Distill Llama 70B	42.4 (20.0)	40.8 (22.0)	45.6 (24.0)	44.2 (18.0)
	deepseek-chat (Deepseek-V3)	58.0 (46.0)	55.6 (38.0)	56.0 (36.0)	58.8 (36.0)
	deepseek-reasoner (Deepseek-R1)	66.8 (48.0)	70.2 (54.0)	72.4 (48.0)	72.8 (56.0)
	gemini-2.5-flash-preview-04-17	66.0 (34.0)	65.2 (38.0)	69.2 (44.0)	72.0 (58.0)
	OpenAI o3	79.6 (66.0)	79.6 (64.0)	82.8 (66.0)	81.2 (68.0)
Logic	DeepSeek R1 Distill Qwen 1.5B	4.0 (4.0)	4.0 (6.0)	6.8 (6.0)	3.6 (4.1)
	DeepSeek R1 Distill Qwen 14B	22.0 (16.0)	23.6 (16.0)	27.2 (22.0)	26.0 (32.0)
	DeepSeek R1 Distill Llama 70B	24.4 (16.0)	24.4 (14.0)	26.0 (20.0)	29.2 (26.0)
	deepseek-chat (Deepseek-V3)	37.8 (30.6)	40.8 (28.0)	41.6 (22.0)	41.4 (24.5)
	deepseek-reasoner (Deepseek-R1)	44.6 (26.0)	45.4 (32)	49.4 (32.7)	50.6 (40.0)
	gemini-2.5-flash-preview-04-17	49.2 (36.0)	51.2 (34.0)	54.0 (42.0)	53.6 (38.0)
	OpenAI o3	68.4 (50.0)	71.2 (54.0)	70.0 (52.0)	74.4 (54.0)

Models struggle at 50 most difficult problems

Some insights that are necessary for solving the problems remain elusive such that models benefit from hints that impart these insights.

Creative Solution Rate:
Human vs Frontier Reasoning Models



Key takeaway

LLMs are sometimes able to identify creative solutions to difficult brainteasers.

But there also remain cases where LLMs fall back to brute-force methods even when a more efficient creative solution exists.

Two levers for increasing creative solution rate

Scaling: while larger models appear naturally less reliant on brute-force

Prompting remains a powerful mechanism for aligning model behavior with human-like problem-solving expectations.

Other Takeaways

LLMs struggle to correct solutions based on gold solutions.

Translating from verbal narratives into mathematical-style problem statements provides modest gains in performance.

Strongest reasoning models are able to reliably break down the solutions into insightful steps and models are capable of using high-level steps to generate correct solutions for the hardest problems where it fails at.

Thank you!

Data & Code