Member-only story

Apple Just Pulled the Plug on the AI Hype. Here's What Their Shocking Study Found

New research reveals that today's "reasoning" models aren't thinking at all. They're just sophisticated pattern-matchers that completely break down when things get tough

Rohit Kumar Thakur 6 min read · Jun 8, 2025

Photo by Medhat Dawoud on Unsplash

We're living in an era of incredible AI hype. Every week, a new model is announced that promises to "reason," "think," and "plan" better than the last. We hear about OpenAI's o1 o3 o4, Anthropic's "thinking" Claude models, and Google's gemini frontier systems, all pushing us closer to the holy grail of Artificial General Intelligence (AGI). The narrative is clear: AI is learning to think.

But what if it's all just an illusion?

What if these multi-billion dollar models, promoted as the next step in cognitive evolution, are actually just running a more advanced version of autocomplete?

That's the bombshell conclusion from a quiet, systematic study published by a team of researchers at Apple. They didn't rely on hype or flashy demos. Instead, they put these so-called "Large Reasoning Models" (LRMs) to the test in a controlled environment, and what they found shatters the entire narrative.

In this article, I'm going to break down their findings for you, without the dense academic jargon. Because what they discovered isn't just an incremental finding.. it's a fundamental reality check for the entire AI industry.

Why We've Been Fooled by AI "Reasoning"

First, you have to ask: how do we even test if an AI can "reason"?

Usually, companies point to benchmarks like complex math problems (MATH-500) or coding challenges. And sure, models like Claude 3.7 and DeepSeek-R1 are getting better at these. But the Apple researchers point out a massive flaw in this approach: **data contamination.**

In simple terms, these models have been trained on a huge chunk of the internet. It's highly likely they've already seen the answers to these famous problems, or at least very similar versions, during their training.

Think of it like this: if you give a student a math test but they've already memorized the answer key, are they a genius? Or just good at memorizing?

This is why the researchers threw out the standard benchmarks. Instead, they built a more rigorous proving ground.

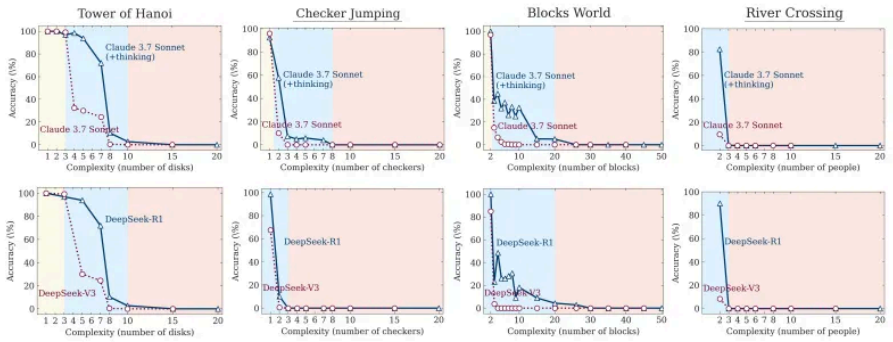
The AI Proving Ground: Puzzles, Not Problems

To truly test reasoning, you need a task that is:

1. **Controllable:** You can make it slightly harder or easier.
2. **Uncontaminated:** The model has almost certainly never seen the exact solution.
3. **Logical:** It follows clear, unbreakable rules.

So, the researchers turned to classic logic puzzles: **Tower of Hanoi, Blocks World, River Crossing, and Checker Jumping.**

These puzzles are perfect. You can't "fudge" the answer. Either you follow the rules and solve it, or you don't. By simply increasing the number of disks in Tower of Hanoi or blocks in Blocks World, they could precisely crank up the complexity and watch how the AI responded.



This is where the illusion of thinking began to crumble.

The Shocking Discovery: AI Hits a Brick Wall

When they ran the tests, a clear and disturbing pattern emerged. The performance of these advanced reasoning models didn't just decline as problems got harder — it fell off a cliff.

The researchers identified three distinct regimes of performance:

- **Low-Complexity Tasks:** Here's the first surprise. On simple puzzles, standard models (like the regular Claude 3.7 Sonnet) actually *outperformed* their "thinking" counterparts. They were faster, more accurate, and used far fewer computational resources. The extra "thinking" was just inefficient overhead.
- **Medium-Complexity Tasks:** This is the sweet spot where the reasoning models finally showed an advantage. The extra "thinking" time and chain-of-thought processing helped them solve problems that the standard models couldn't. This is the zone

that AI companies love to demo. It looks like real progress.

- **High-Complexity Tasks:** And this is where it all goes wrong. Beyond a certain complexity threshold, **both model types experienced a complete and total collapse.** Their accuracy plummeted to zero. Not 10%. Not 5%. Zero.

This isn't a graceful degradation. It's a fundamental failure. The models that could solve a 7-disk Tower of Hanoi puzzle were utterly incapable of solving a 10-disk one, even though the underlying logic is identical. This finding alone destroys the narrative that these models have developed generalizable reasoning skills.

Even Weirder: When the Going Gets Tough, AI Gives Up

This is where the study gets truly bizarre. You would assume that when a problem gets harder, a "thinking" model would.. well, *think harder*. It would use more of its allocated processing power and token budget to work through the more complex steps.

But the Apple researchers found the exact opposite.

As the puzzles approached the complexity level where the models would fail, they started to use *fewer* tokens for their "thinking" process.

Let that sink in.

Faced with a harder challenge, the AI's reasoning effort *decreased*. It's like a marathon runner who, upon seeing a steep hill at mile 20, decides to start walking slower instead of digging deeper, even though they have plenty of energy left. It's a counter-intuitive and deeply illogical behavior that suggests the model "knows" it's out of its depth and simply gives up.

This reveals a fundamental scaling limitation. These models aren't just failing because the problems are too hard; their internal mechanisms actively disengage when faced with true complexity.

Inside the AI's "Mind": A Tale of Overthinking and Underthinking

The researchers didn't stop at just measuring final accuracy. They went deeper, analyzing the "thought" process of the models step-by-step to see *how* they were failing.

What they found was a story of profound inefficiency.

- **On easy problems, models "overthink."** They would often find the correct solution very early in their thought process. But instead of stopping and giving the answer, they would continue to explore dozens of incorrect paths, wasting massive amounts of computation. It's like finding your keys and then spending another 20 minutes searching the rest of the house "just in case."

- **On hard problems, models "underthink."** This is the flip side of the collapse. When the complexity was high, the models failed to find *any* correct intermediate solutions. Their thought process was just a jumble of failed attempts from the very beginning. They never even got on the right track.

Both overthinking on easy tasks and underthinking on hard ones reveal a core weakness: the models lack robust self-correction and an efficient search strategy. They are either spinning their wheels or getting completely lost.

The Final Nail in the Coffin: The "Cheat Sheet" Test

If there was any lingering doubt about whether these models were truly reasoning, the researchers designed one final, damning experiment.

They took the Tower of Hanoi puzzle: a task with a well-known, recursive algorithm and literally gave the AI the answer key. They provided the model with a perfect, step-by-step pseudocode algorithm for solving the puzzle. The model's only job was to *execute* the instructions. It didn't have to invent a strategy; it just had to follow the recipe.

The result?

The models still failed at the exact same complexity level.

This is the most crucial finding in the entire paper. It proves that the limitation isn't in problem-solving or high-level planning. The limitation is in the model's inability to consistently follow a chain of logical steps. If an AI can't even follow explicit instructions for a simple, rule-based task, then it is not "reasoning" in any meaningful human sense.

It's just matching patterns. And when the pattern gets too long or complex, the whole system breaks.

So, What Are We Actually Witnessing?

The Apple study, titled "The Illusion of Thinking," forces us to confront an uncomfortable truth. The "reasoning" we're seeing in today's most advanced AI models is not a budding form of general intelligence.

It is an incredibly sophisticated form of pattern matching, so advanced that it can *mimic* the output of human reasoning for a narrow band of problems. But when tested in a controlled way, its fragility is exposed. It lacks the robust, generalizable, and symbolic logic that underpins true intelligence.

The bottom line from Apple's research is stark: we're not witnessing the birth of AI reasoning. We're seeing the limits of very expensive autocomplete that breaks when it matters most.

The AGI timeline didn't just get a reality check. It might have been reset entirely.

So the next time you hear about a new AI that can "reason," ask yourself: Can it solve a simple puzzle it's never seen before? Or is it just running the most expensive and convincing magic trick in history?

Apple AI AGI Artificial Intelligence OpenAI

Written by Rohit Kumar Thakur

3.3K followers · 9 following

I write about AI, Tech, Startup and Code

Follow

Responses (269)

Harry Baya

What are your thoughts?

Dimitre Novatchev
Jun 11 (edited)

Thank you for this informative article.

Inspired by it, I asked CoPilot if it were up to the task of solving Sudoku puzzles. I provided the initial board using chess-like coordinate system (a - i, 1-9)

It replied confidently that it could, and even... more

1K 6 replies Reply

Andrei Dascalu
Jun 10

A good sum up of just about every technical insider has been saying all along. I'm not sure about the "study" needed since the people working on these models have said as much when discussing what they can really do.

352 3 replies Reply

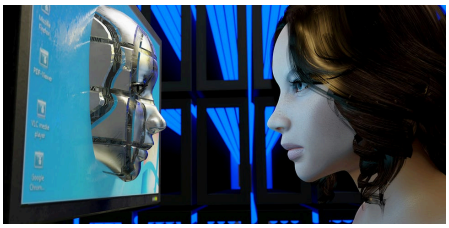
Melody Na
Jun 10

'It is an incredibly sophisticated form of pattern matching, so advanced that it can mimic the output of human reasoning for a narrow band of problems. But when tested in a controlled way, its fragility is exposed. It lacks the robust...
more

291 11 replies Reply

See all responses

More from Rohit Kumar Thakur



Rohit Kumar Thakur

Rohit Kumar Thakur

We Just Discovered a Trojan Horse in AI, And...

A model trained on nothing but numbers can learn to be...

Jul 30 4.5K 152

The Era of the AI Black Box Is Officially Dead....

For years, we guessed what an AI was thinking. A new paper...

Aug 2 2.3K 95

Rohit Kumar Thakur

Google DeepMind Just Dropped a 'Transforme...

It gets 2x inference speed, slashes memory usage in half,...

Jul 21 1.4K 29

Rohit Kumar Thakur

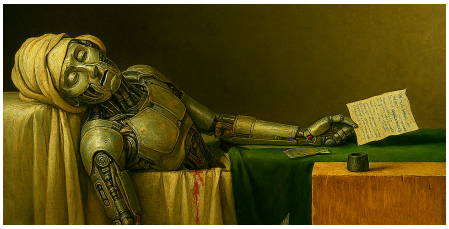
It's Game Over: The Real "AlphaGo Moment" Jus...

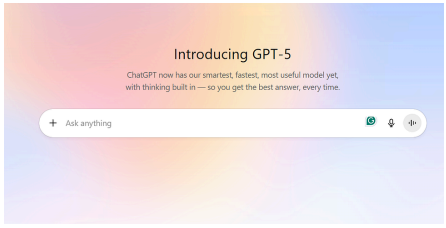
A new paper just revealed an AI that invents its own technology...

Jul 26 2.2K 65

See all from Rohit Kumar Thakur

Recommended from Medium





Jordan Gibbs

OpenAI Just Gutted ChatGPT.

GPT-5 is nothing but a poorly disguised downgrade.

5d ago 2.4K 122

Alberto Romero

GPT-5 Is Here: There's Only One Feature Wort...

My short review of OpenAI's new flagship model

6d ago 1.9K 76

In Predict by iswarya writes

Future-Proof Careers in the Age of AI: What You...

What if I told you that by this time next year, you could land a...

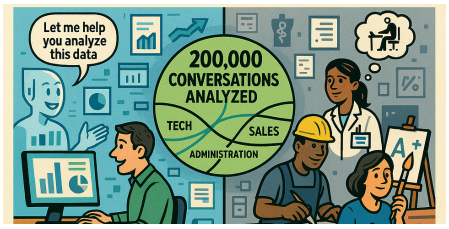
Jul 29 2.1K 107

In Long. Sweet. V... by Ossai Chin...

I'll Instantly Know You Used Chat Gpt If I See...

Trust me you're not as slick as you think

May 16 20K 1222



In Write A Catal... by MKWritesh...

Microsoft Study: 40 Jobs AI Will Replace vs 40...

200,000 real workplace conversations reveal which job...

Jul 29 764 23

In AI Mind by Mr Tony Momoh

5 Industries AI Will Completely Take Over ...

It's Already Started

Jun 10 8.8K 389

See more recommendations

All your favorite parts of Medium are now in one sidebar for easy access.

Home

Library

Profile

Stories

Stats

Following >

Discover more writers and publications to follow.

See suggestions