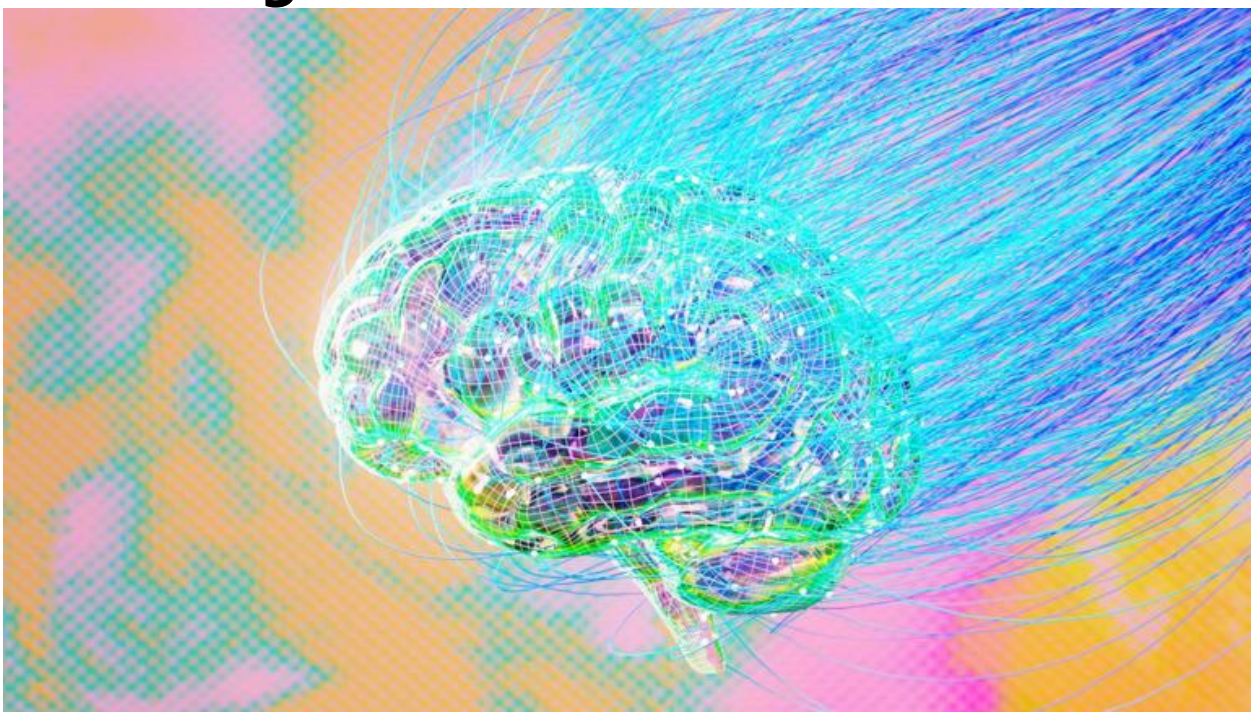
# Scientists just developed a new AI modeled on the human brain — it's outperforming LLMs like ChatGPT at reasoning tasks



Scientists just developed a new AI modeled on the human brain — it's outperforming LLMs like ChatGPT at reasoning tasks

Scientists have developed a new type of artificial intelligence (AI) model that can reason differently from most large language models (LLMs) like ChatGPT, resulting in much better performance in key benchmarks.

The new reasoning AI, called a hierarchical reasoning model (HRM), is inspired by the hierarchical and multi-timescale processing in the human brain — the way different brain regions integrate information over varying durations (from milliseconds to minutes).

Scientists at Sapient, an AI company in Singapore, say this reasoning model can achieve better performance and can work more efficiently. This is thanks to the model requiring fewer parameters and training examples.

The HRM model has 27 million parameters while using 1,000 training samples, the scientists said in a study uploaded June 26 to the preprint arXiv database (which has yet to be peer-reviewed). In comparison, most advanced LLMs have billions or even trillions of parameters. Although an exact figure has not been made public, some estimates suggest that the newly released GPT-5 has between 3 trillion and 5 trillion parameters.

## A new way of thinking for AI

When the researchers tested HRM in the ARC-AGI benchmark — a notoriously tough examination that aims to test how close models are to achieving artificial general intelligence (AGI) — the system achieved impressive results, according to the study.

HRM scored 40.3% in ARC-AGI-1, compared with 34.5% for OpenAI's o3-mini-high, 21.2% for Anthropic's Claude 3.7 and 15.8% for Deepseek R1. In the tougher ARC-AGI-2 test, HRM scored 5% versus o3-mini-high's 3%, Deepseek R1's 1.3% and Claude 3.7's 0.9%.

Most advanced LLMs use chain-of-thought (CoT) reasoning, in which a complex problem is broken down into multiple, much simpler intermediate steps that are expressed in natural language.

It emulates the human thought process by breaking down elaborate problems into digestible chunks.

But the Sapient scientists argue in the study that CoT has key shortcomings — namely "brittle task decomposition, extensive data requirements, and high latency."

Instead, HRM executes sequential reasoning tasks in a single forward pass, without any explicit supervision of the intermediate steps, through two modules. One high-level module is responsible for slow, abstract planning, while a low-level module handles rapid and detailed computations. This is similar to the way in which the human brain processes information in different regions.

It operates by applying iterative refinement — a computing technique that improves the accuracy of a solution by repeatedly refining an initial approximation — over several short bursts of "thinking." Each burst considers whether the process of thinking should continue or be submitted as a "final" answer to the initial prompt.

But the Sapient scientists argue in the study that CoT has key shortcomings — namely "brittle task decomposition, extensive data requirements, and high latency."

Instead, HRM executes sequential reasoning tasks in a single forward pass, without any explicit supervision of the intermediate steps, through two modules. One high-level module is responsible for slow, abstract planning, while a low-level module handles rapid and detailed computations. This is similar to the way in which the human brain processes information in different regions.

It operates by applying iterative refinement — a computing technique that improves the accuracy of a solution by repeatedly refining an initial approximation — over several short bursts of "thinking." Each burst considers whether the process of thinking should continue or be submitted as a "final" answer to the initial prompt.