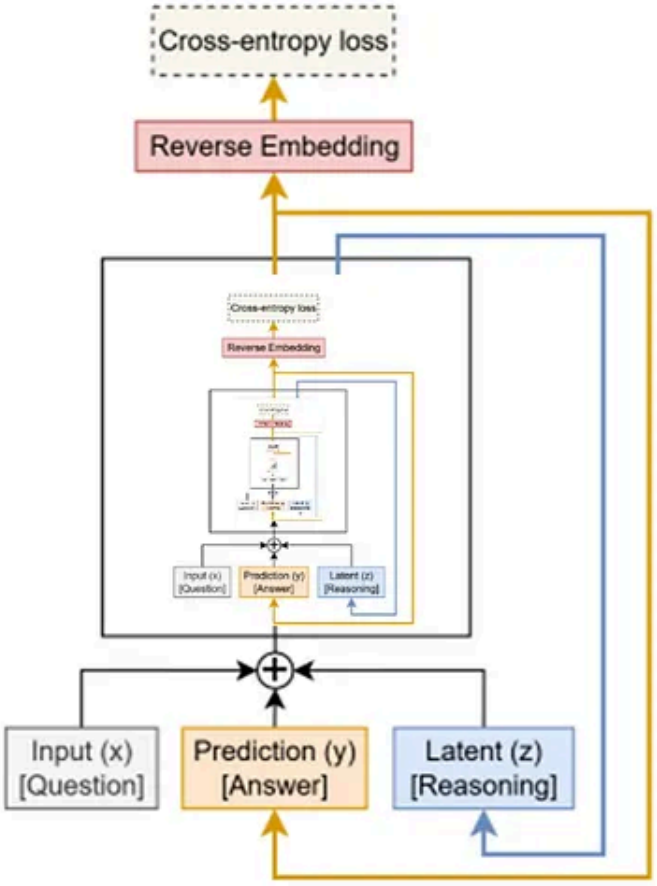
Member-only story

Silicon Valley Is Obsessed With the Wrong AI

But there are interesting alternatives, like Tiny Recursion Models (TRMs)

21 min read · 3 days ago

Alberto Romero Follow



The schematic view of a Tiny Recursion Model but it's actually a recursion joke

I. The AI industry is hanging by a thread

In August 2025, I wrote the excerpt below about the AI bubble in an article that you guys loved:

> A bubble, it is often believed, spawns out of thin air; there's nothing of value hiding inside but a void made out of lies and vaporware waiting to fall upon the world like a heavy rain. But that's rarely the case. Bubbles are simply an unhealthy extension of the real value lying at the center. There is a "kernel of truth," as OpenAI CEO Sam Altman, the modern "Bubble Man" par excellence, told The Verge.

I went on to explain what happens when that kernel of truth is buried under promises and overinvestment and whatnot, the usual story. *This* is not the usual story. Bubbles are defined by how far from that kernel of truth they can be expanded before bursting; they are never defined by how brittle that kernel of truth is.

In another article from September 2025, I compared the AI bubble and the railroad mania: the railroad's contribution to society was always clear, before there was any mania about them! I wrote:

> It's not hard to understand excess bubbles like the railway mania in the mid-1800s, or the canal mania in the late 1700s, or the tulip mania in the mid-1600s (the speculation is about scale: "We may not need as much of X, but someone will get richer"). It is harder to understand *betting* bubbles (the speculation is about *viability*: "X may or may not work, so let's gamble"). AI is both.

Ok, this is terrible news!

But although I did a lot of explaining in both articles, I never got to answer a fundamental question: what is it about AI's kernel of truth that makes it a bet?

I babbled about large language models (LLMs) being unreliable, which is true, and also about the political, sociocultural, geopolitical, and economic dimensions of the bet. But I never deigned to find a concrete technical example that might anchor the argument in something tangible, a visible crack in that kernel of truth we are taking for granted.

Then, in "The Neuron That Wanted to Be God," which I published in October, I temporarily became a historian of AI and found an eighty-year-old dark secret that's not so secret (if you only learn one thing about the history of the field, let it be this): the edifice of modern AI is erected on top of a blatant simplification of the elemental building block — the neuron — that gives rise to intelligence in humans. But still, although you guys loved that one as well, it was perhaps too abstract and maybe not at all convincing. Is it that important if the artificial neuron that

underlies layers and layers of abstractions inside ChatGPT resembles nothing the biological neuron that underlies your brain? The only valid response is *I don't know*; **one does not break a kernel of truth with an uncertain alternative!

What I bring you today is an in-depth exploration of a timely topic that reveals you don't need to go search in the archives from the last century to find a good reason to 1) doubt the technical core of the AI industry — that scaling LLMs is what you need to reach the infinite — and 2) doubt the financial core: maybe $1 trillion in investment to build datacenters to train and serve gigantic and expensive LLMs is *unnecessary*. I believe that Silicon Valley has grown too obsessed (and complacent) with the status quo. I believe, and so do many others, that there's a big incentive to discover an AI breakthrough that, on the one hand, displays competence where LLMs are brittle and unreliable, and, on the other, doesn't require gargantuan levels of spending.

Today, I choose to open a crack in that kernel of truth (I urge you to read until the end because, spoiler, there are plot twists). For that, we have to travel in space and time; we need to go back to July 2025 in Singapore, to the headquarters of a barely known AI research lab founded by graduates from Tsinghua University, the top technical university in China, which is generating AI talent at a pace that Western universities, combined, can't match.

It is at Sapient Intelligence that our story begins, and it begins with a crazy idea: A brain-inspired model, a rebellion against LLMs pre-training on the entire web, and a sound victory over much wealthier AI labs, sleeping unaware that the ground is shaking, at the other end of the world.

This story begins with the *Hierarchical Reasoning Model*.

II. The breakthrough: Hierarchical Reasoning Model

Guan Wang et al. published a research paper on arXiv in July 2025, accompanied by a Twitter thread that instantly grabbed the attention of the AI community:

> Inspired by brain's hierarchical processing, HRM delivers unprecedented reasoning power on complex tasks like ARC-AGI and expert-level Sudoku using just 1k examples, no pretraining or CoT! Unlock next AI breakthrough with neuroscience.

A few things should be immediately put in quarantine: bioinspiration? 1,000 examples? No pre-training? No CoT? Beating reasoning LLMs on ARC-AGI? Each

element of this list of rhetorical questions is, potentially, paradigm-shifting — how could AI labs have missed something like this? This is either huge, I thought, or a bad joke. But given that I'm writing this and you're not laughing, I guess we already know.

This will get a bit technical, but I promise to keep it as clear as possible. The main motivation for Wang et al. to explore what they call the Hierarchical Reasoning Model (HRM) is that LLMs require chain of thought (CoT) to do reasoning. This is expensive, data-intensive, and high-latency (slow). CoT underlies every commercial LLM worth using nowadays. You can't get ChatGPT to solve math and coding problems without it.

In most cases, you can't access the CoT of the model because companies hide it, but, for instance, DeepSeek-R1 was beloved precisely because they kept it visible (visible CoTs had unforeseen consequences). You'd see the R1 talking to itself through the problems before answering, going back and forth, backtracking, going on a tangent, and so on, more or less like humans do.

Wang et al. consider CoT *a bad approach*, one that only exists because researchers at AI labs in the West didn't bother to look inside their own brains for inspiration.

HRMs, in contrast, were "Inspired by the hierarchical and multi-timescale processing in the human brain." I guess I'm not the only one interested in revisiting the intermediate layers that underlie LLMs to make them resemble the human brain! To some degree, good ol' recurrent and convolutional neural networks (RNNs and CNNs; many of you may not have heard of them at all) are like that. But the famous transformer (Vaswani et al., 2017) basis of all modern LLMs, removed both recurrence and convolution in favor of the attention mechanism, hence the title "Attention is all you need." Interestingly, the attention mechanism itself is loosely inspired by the brain, but that's as far as modern LLMs resemble our cognitive engine (which is to say, *not much*).

Let's look into those quarantined items above in more detail. A dataset of 1,000 examples is virtually insignificant measured against the scale of LLMs, which are trained on dozens of trillions of tokens (words). But of course, this is not an apples-to-apples comparison because Wang et al. didn't test HRM on language tasks; they focused on puzzles like Sudokus and, something I'm particularly interested in, the ARC-AGI benchmark. (This doesn't mean that HRM is irrelevant because it doesn't

know language, nor that it will surely beat LLMs once it learns, because it beat them at ARC-AGI; we can make no assertions about that, so let's keep the comparisons as low-key as possible.)

So, HRMs display impressive training data efficiency and also reject the pre-training paradigm. You may have heard about pre-training in the context of the scaling laws: AI labs increase computing power and dataset size to improve AI models through pre-training — aka eating all the data on the internet — but once the scaling laws began to show diminishing returns late in 2024, companies shifted to post-training (reinforcement learning on reasoning data) and test-time compute (allowing the models to think before responding). Nowadays, OpenAI, Google, Meta, etc., are busy scaling post-training and have set aside (not completely!) scaling pre-training. But none of them, not even for a moment, have considered killing pre-training altogether.

That makes HRM a *different paradigm*. Don't confuse this with HRMs being a paradigm shift that *improves* over the existing one. HRMs are not necessarily better — not even in theory or principle! I've written about the possibility that AI could learn without eating the internet in the context of DeepSeek-R1 Zero here and here. My point is that just like AlphaGo Zero learned to play Go purely through self-play — without human guidance and without looking at the archive of human games — and promptly surpassed AlphaGo, which trained on human games, the same could happen with language and reasoning. I love that idea because, as AI pioneer Richard Sutton says, humans don't learn to do basic stuff by reading the entire corpus of books on "how to do basic stuff." We experiment and do trial-and-error (and enjoy a rather helpful genetic endowment).

In any case, I don't want to get ahead of myself: although it's an appealing possibility, no one knows whether it's possible to achieve language mastery without first eating the entire internet.

Another important detail about HRM before going into the results of the study (we'll get there!) is that it is an extremely small model, "just" 27 million parameters. I wonder if the quotation marks are even needed, considering that GPT-4, for instance, was estimated to have 1.8 *trillion* parameters. That's 1.8 million million parameters (to make the comparison clearer), or 100,000x the size of HRM. I said I didn't want to do direct comparisons because, again, this is not apples-to-apples, but this is a fundamental difference between LLMs and basically *any other neural*

network that may attempt to dethrone them as the cornerstone of the AI edifice. Enough words, let's see the chart:

HRM surpasses state-of-the-art CoT models on ARC-AGI, Sudoku-Extreme, and Maze-Hard. (Source)

> HRM, trained from scratch with only the official dataset (~1000 examples), with only 27M parameters and a 30x30 grid context (900 tokens), achieves a performance of 40.3%, which substantially surpasses leading CoT-based models like o3-mini-high (34.5%) and Claude 3.7 8K context (21.2%), despite their considerably larger parameter sizes and context lengths.

The first two from the left are ARC-AGI 1 and 2. HRM outperforms DeepSeek-R1, Claude Sonnet 3.7, and o3-mini-high, all of which had feasted on humanity's data before attempting this challenge. How is this possible? Even if it's not an apples-to-apples comparison in the broader sense because HRM hasn't learned language, these results on ARC-AGI 1 and 2 *are a fair contest* in the sense that solving them should be trivial for powerful models! And yet, HRM — brain-inspired, no pre-training, no CoT, teeny tiny — wins.

Let's see what Wang et al. offer as an explanation of this unexpected success.

III. Face-to-face: HRMs versus LLMs

The authors make a concrete criticism about the architecture of LLMs, which is the first point of discrepancy with HRMs: "Despite the remarkable success of large language models, their core architecture is paradoxically shallow." With *paradoxically shallow*, they mean that although transformers look complex, the theoretical bounds on their computational power prevent them from doing "complex algorithmic reasoning." They argue this is the reason behind the inability

of powerful LLMs to beat puzzles like ARC-AGI. (There's no need to understand the jargon-y details here, which I'm mostly omitting for the sake of clarity, although you can always read the paper and the literature!). HRM presumably solves this shallowness in transformers by using two coupled modules: a slow, high-level module for planning, and a fast, low-level module for execution.

But AI labs don't play under the same pressures or preferences: they need AI models to be *commercially viable* first, and only then capable of achieving great feats of reasoning. This is the logic of the real world: you may have the blueprint to build a hypersonic plane, but if it costs 100x more than a plane that goes half as fast, you'd choose the latter every time.

So, instead of over-complicating the architecture with bioinspiration (commercially inviable for complex reasons; you're not going to re-engineer evolution's masterpiece), they improve LLMs' reasoning capabilities in some other way: they use chains of thought. They allow the LLM to "think out loud" the intermediate steps of the problem using human language (explicit words) instead of thinking internally. By doing this, they displace the bottleneck from "transformers can't reason and that's bad" to "transformers can reason by leveraging a CoT patch and that's bad." There's still some badness, but it's at least of the commercially viable sort! However, they can't rest just yet; they solved the LLM reasoning, but now they have to deal with the patch.

To put it in plain terms: LLMs are the manifestation of a trade-off that prioritizes engineering simplicity over scientific elegance. They look "uglier" because AI companies need to circumvent their inability to reason internally, but, in exchange, they are parallelizable and scalable so that you, demanding customer, can use ChatGPT at all. HRM, instead, avoids this "patch" by resembling the only thing in nature that does reasoning for a living, the human brain. HRMs keep the reasoning process inward, so that it takes place in the "latent space," which is analogous to the "space of thoughts." Whereas LLMs need to write down their thoughts, HRMs don't.

Important figures in the field have talked about this. One prominent example is Meta FAIR's Chief AI Scientist, Yann LeCun, who insists that it makes no sense to force AI systems to reason like humans when reasoning in the continuous latent space allows for much richer computations and more bandwidth than in the sequential, discrete space of words. As he says, "It is intuitively obvious that reasoning in continuous embedding space is dramatically more powerful than

reasoning in discrete token space." Here's a paper referenced by Wang et al. of work done at Meta FAIR (by Hao et al., 2024). Here's the paper referenced by LeCun in that tweet, also at FAIR (by Zhu et al., 2025).

I agree with this highly intuitive criticism of LLMs and the paradigm of CoT: it's an unserious argument that we should stick to LLMs writing down their thoughts only because we can't find a better architecture that allows them to do it at the level of thoughts and, at the same time, be commercially viable. Reasoning at the level of words is nevertheless useful — in part, humans write to think — but we're not restricted to this mechanism and surely it's not the main one, for despite the prevalent belief that language is a tool for thought rather than communication, recent research shows that it is indeed, "a tool for communication more than thought."

So HRMs think inwardly and are brain-like, but what does that look like in practice? How do they reason in concrete terms? They reason by doing "thinking bursts," as the ARC-AGI team calls them. The HRM will propose solutions to the tasks (e.g., an ARC-AGI 1 puzzle) and then iterate over them internally, concluding at each step that it either needs to refine the solution or stop. It's within this iterative refinement process that the slow planner and the fast executor work together to devise a new answer. (This is an example of test-time training; HRMs save compute costs during the training phase because they learn to solve the task at inference/test time.)

Ok, but let's be frank, who cares about HRMs when LLMs are in vogue, and everywhere underlie the most powerful AI systems in the world, namely, ChatGPT, Gemini, Claude, Grok, Llama, DeepSeek…? Well, here's my non-exhaustive list — HRMs are great news for:

1. People who think scaling LLMs is a brute-force approach that can't possibly work if we don't go through a few algorithmic breakthroughs in the meantime.
2. People who dislike the idea of forcing LLMs to reason "out loud" and favor a "latent/token space" reasoning instead; all human writing is thinking, but not all human thinking is writing.
3. People who are angry that AI labs are stealing everyone's data to train compute-intensive and data-intensive LLMs that are infamously unreliable.

4. People who think that AI systems should take more inspiration from biology and the human brain than they currently do.
5. People who think there's a financial AI bubble because it makes no sense to spend trillions of dollars on datacenter infrastructure before revisiting the fundamentals.
6. People who love the ARC-AGI benchmark as among the few AI evaluations that, being easy for humans, constantly trip up the best AI models.

But, as fate would have it — plot twist, I warned you! — it turns out that HRMs are not as great as the paper (or this section) suggests.

First, the authors were accused of training on test data (data leakage), but, fortunately, this was confirmed false (the paper and HRMs are legit), so that's not the reason (in case you had read about this and didn't know how it was resolved). Second, there are various shortcomings that the authors openly acknowledge: scaling is always hard and may not even be possible with HRMs, training and inference are coupled, and that makes them expensive, HRM is not a "general-purpose" model (i.e., it can only solve tasks it's previously seen). This may prove fatal for the HRM approach.

Finally, there's one other flaw that the authors apparently missed, and it is, in my opinion, the critical one. The main proposal of HRMs — namely, that they can reason in the latent space vs token space and achieve greater "computational depth" than LLMs without pre-training or CoTs because they're inspired by the brain's hierarchical and multi-timescale processing — is *mostly irrelevant!*

The authors unintentionally buried the lede under a lot of wrong conclusions about bioinspiration and such. The actual breakthrough of this study and the main driver of performance on ARC-AGI comes, as the ARC-AGI team wrote, from an "unexpected source." Let's see what's going on.

IV. The hidden drivers of HRM's performance

First of all, as I said, HRM's performance on ARC-AGI is confirmed, as well as the legitimacy of the paper. François Chollet, co-founder of ARC Prize and creator of ARC-AGI, said this on Twitter:

> ARC-AGI-1: 32% — Though not state of the art, this is impressive for such a small model.
>
> ARC-AGI-2: 2% — While scores >0% show some signal, we do not consider this material progress on ARC-AGI-2.

ARC-AGI-1 Leaderboard with HRM performance over cost per task. (Source)

The ARC-AGI team found that the outer loop (what Wang et al. call "Recurrent connectivity") is the actual reason why HRMs beat LLMs at ARC-AGI. They ran ablation experiments (important!) to determine whether there were uninteresting elements or even unintended handicaps in the construction of the HRM and found this:

> The HRM model architecture itself (the centerpiece of the paper) is not an important factor [and instead] the outer refinement loop (barely mentioned in the paper) is the main driver of performance.

This means that the one element that carried the success has nothing to do with the human brain. They also say that "The 'hierarchical' architecture had minimal performance impact when compared to a similarly sized transformer." Bummer! Transformer-based LLMs win again!

"Pass@2 performance over varying number of training and inference refinement loops. Iterating over data by refining it has a strong impact, as the jump from 1 (no refinement) to 2 (1 refinement) shows." (Source)

The outer loop "feeds the model output *back* into itself, allowing the model to iteratively refine its predictions." So, it's not the actual task-solving of the coupled modules (slow planner + fast executor) or the specifics of the architecture that provide performance gains, but the fact that the solutions are *fed back into the model*. (The number of loops you do impacts performance a lot, as you can see in the chart above.)

The ARC-AGI team compares this stripped-down HRM with the Universal Transformer, a version of the original transformer conceived by researchers at Google and DeepMind (DehGhani et al., 2018) that uses a recurrent loop to improve generalization.

If you're wondering why modern LLMs are not based on the Universal Transformer (they aren't!), it is because recurrence breaks the parallelism that makes Transformers trainable on modern hardware. Every recurrent step has to wait for the previous one, which is fundamentally at odds with GPU/TPU architectures designed for batched, simultaneous matrix operations. In contrast, the standard Transformer — though wasteful in some other ways that you know well — maps perfectly onto today's compute stack: it's easy to parallelize, pipeline, and scale across dozens of thousands of accelerators.

So, in practice, AI labs accepted *another* trade-off: algorithmic inefficiency that scales with compute (the vanilla Transformer) over algorithmic efficiency that doesn't scale (the Universal Transformer). Even if HRMs worked on language domains, AI labs would accept the same trade-off.

All in all, HRM was a promising idea (although orders of magnitude less developed in both theory and practice than the LLMs that power ChatGPT, Gemini, Claude, etc.), but the authors failed to nail down the actual breakthrough.

It was Alexia Jolicoeur-Martineau, a researcher at Samsung SAIL, who picked up where the team at Sapient left off. From Singapore, we travel to Montreal, Canada. Alexia refined their approach, tested it, uncovered its weak points, stripped away what wasn't needed, isolated the refinement loop, and presented her findings in an October paper: "Less is More: Recursive Reasoning with Tiny Networks."

V. Refining the loop: Tiny Recursive Model (TRM)

The key insight Alexia had is that although HRMs might work to surpass LLMs on ARC-AGI and other puzzles, being smaller models (27M parameters vs. hundreds of billions) and trained on a small dataset (~1,000 examples) without pre-training, they "may be suboptimal." That is, indeed, what the ARC-AGI team independently figured out (Alexia came up with TRMs inspired by their findings). Here's the main contribution of the TRM research paper:

> We propose Tiny Recursive Model (TRM), a much simpler recursive reasoning approach that achieves significantly higher generalization than HRM, while using a single tiny network with only 2 layers. With only 7M parameters, TRM obtains 45% test-accuracy on ARC-AGI1 and 8% on ARC-AGI-2, higher than most LLMs (e.g., Deepseek R1, o3-mini, Gemini 2.5 Pro) with less than 0.01% of the parameters.

Alexia's approach: Trim the excess even if it sounds good (the bio-inspired hierarchical architecture, i.e., the slow planner and the fast executor are gone, and other technical details that are irrelevant for our purposes here but that Alexia skillfully ablated away), improve the weaknesses (simplify the recursive process to make TRMs more generalizable), and package it in a way that others can easily build on and reproduce (compare 7 million parameters vs 1.8 trillion in GPT-4).

The basic idea of TRMs is pretty similar but much simpler than HRMs. TRM searches for the solution to a given problem by iterating and improving over its

previous reasoning and previous answers, but without all the unnecessary paraphernalia. Here's a high-level explanation and a picture:

Tiny Recursion Model (Source)

Every step for a given number of steps, the TRM takes the question-answer (x, y) pairs as well as the current reasoning (z, latent) and uses them to update the reasoning for the next step (basically, the model needs to remember what it reasoned and answered before to improve). That's what Alexia calls "recursive reasoning." (Note that during this part of the process, the answer is not updated!) This is all pretty intuitive because humans also do this: if you want to improve your answer to a math problem, you need to know what answer you gave that was wrong and also what reasoning led you to the wrong answer. Pretty straightforward stuff.

So the TRM updates the reasoning a few times and, taking the last reasoning, z, paired with the previous answer, y, then proposes a new answer. Those two parts

comprise the entire recursive loop: 1) do recursive reasoning a given number of times, and 2) update and provide the answer.

Importantly, like HRMs, TRMs avoid doing CoT. The reasoning takes place in the latent space (internally) rather than in token space (written out) as LLMs do. Besides the obvious motivation to avoid CoT — thinking aloud constraints reasoning to what language can handle and it's ugly — Alexia notes, as Wang et al. do, that it's extremely expensive computationally and requires high-quality data.

(You wouldn't imagine how much money AI labs spend on hiring contractors to provide them with labeled "reasoning data," i.e., question-answer pairs for hard problems in, say, coding, math, etc., to do reinforcement learning on it, only for AI heavyweight Andrej Karpathy to say that that's a "terrible" approach, that the models are merely "sucking supervision through a straw"!)

However, even if TRM is superior to HRM in both simplicity and generalizability, it's an open question whether the success on ARC-AGI can be extrapolated to other domains like language and coding, where generative AI can be useful (note that TRM is not a model of the generative kind just yet, but it can be further developed into one). Whatever the case, these scores (below) on ARC-AGI by such a tiny, simple model are, to me, thought-provoking, and to the AI labs, anxiety-inducing:

"% Test accuracy on ARC-AGI Benchmarks (2 tries)" (Source)

A 7-million parameter TRM surpasses some of the top LLMs (Gemini 2.5 Pro, o3-mini-high, Claude 3.7, and DeepSeek-R1) on both ARC-AGI 1 and 2. This is incredible. Alexia considers what is perhaps the most important technical question: Why does deep recursion work better than using a much larger language model? Her main suspect is overfitting (too many parameters that learn the distribution too well and prevent the model from generalizing over to out-of-distribution problems), but, as the paper says, there's "no theory to back this explanation." Gut feeling rules AI!

To avoid adding to the initial controversy that surrounded HRMs back in July, I waited until the ARC-AGI team corroborated the results on Alexia's TRM on their own semi-private ARC-AGI evaluation test. They just did. TRM doesn't surpass all commercial LLMs — not even in terms of price at times (e.g., Grok 4 Thinking costs the same per task and gets 16% on ARC-AGI 2, and Claude Sonnet 4.5 is ~10x cheaper and gets 14%; this is not counting the costs of training, which is where most of the money goes for LLMs) — but it is a genuine breakthrough that the community will keep an eye on.

ARC-AGI 1 leaderboard (Source)

ARC-AGI 2 leaderboard (note the scale is up to 35%; no model — LLM, TRM or otherwise — has solved it!) (Source)

VI. Silicon Valley is obsessed when it should be open-minded

Before closing this exploration, I want to get back to the main distinction between HRMs and TRMs because I think there's an important, albeit nuanced, lesson there.

If you've read "The neuron that wanted to be God," which I mentioned in the beginning, you know that I favor revisiting old assumptions and simplifications that might be problematic for the creation of intelligence of the fluid rather than crystallized kind (to use Chollet's nomenclature), but I don't think "the brain works like this"-type of bioinspiration should be imposed onto AI architectures without justification — HRMs are an example of forcing reality to fit a preconceived model rather than letting the model fit reality.

Alexia writes about why HRM doesn't make sense for this reason:

> The HRM's authors justify the two latent variables and two networks operating at different hierarchies based on biological arguments, which are very far from artificial neural networks. They even try to match HRM to actual brain experiments on mice. While interesting, this sort of explanation makes it incredibly hard to parse out why HRM is designed the way it is. Given the lack of ablation table in their paper, the over-reliance on biological arguments and fixed-point theorems (that are

> not perfectly applicable), it is hard to determine what parts of HRM is helping what and why.

However, she directly agrees with my broader point, which isn't to copy the brain (that makes little sense!) but *not to dismiss* possibilities across the entire ladder — from neuron to network to model to system to product — that might suggest an alternative to standard LLMs and hundreds of billions of dollars in investment, which only big Silicon Valley companies can afford.

Neuro-symbolic approaches are naturally looked down upon because, in the past, overconfidence in symbolism led to two AI winters and a long financial drought for the field. But it's weird that, given how good techies are at abstraction and meta-thinking, they're afraid and resentful of symbolic AI in particular rather than of *overconfidence in general*.

I've kept an eye on ARC-AGI for years because I intuited that LLMs still fail to beat it for this reason: the industry, which owes its very existence to the deep pockets of AI labs and shareholders, had to accept a very delicate trade-off to mature. As I wrote above, they're exchanging inefficiency and expensiveness to counter the risk of infeasibility. LLMs might be working for now because AI labs have enough compute and data to make them work, but one must be blind — or perhaps one's salary/identity depends on it — if one doesn't think the sector has crossed a few lines in terms of energy usage and required investment that should make one, at the very least, wary and suspicious.

TRMs are a breakthrough, primarily, because they're evidence that things are only unfeasible if you don't try to make them feasible in the first place. And that we take for granted things (namely, LLMs) out of habit more than certainty. Many research paths have been abandoned not due to failure but due to interest, funding, and pressures leading elsewhere.

Alexia wrote something to this effect on Twitter; although her main contribution is technical, this socioeconomic remark is, in my opinion, just as important:

> The idea that one must rely on massive foundational models trained for millions of dollars by some big corporation in order to solve hard tasks is a trap. Currently, there is too much focus on exploiting LLMs rather than devising and expanding new lines of direction.

Indeed, sometimes “less is more,” and we might be getting to a point where “more is too much.”

. . .

*Join 40,000 others in **The Algorithmic Bridge**, a blog about AI for the people.*

Artificial Intelligence | AI | Technology | Tech | Large Language Models

Follow

Written by Alberto Romero

49K followers · 137 following

AI & Tech | Weekly AI Newsletter: https://thealgorithmicbridge.substack.com/ | Contact: alber.romgar at gmail dot com

Responses (30)

Bgerby

What are your thoughts?

Frank Deutschmann
2 days ago

This is all somewhat interesting, but overall is just clickbait: currently, the "wrong" AI is producing real results, delivering real, functional, value for real paying customers - which makes this AI very much the "right AI for

right now."

In the... more

28 4 replies Reply

Kevin D Kissell
1 day ago

From the title I had hoped this would shine a light on the fact that there are non-language-model forms of AI that are in fact incredibly powerful and useful in industry and everyday life, but which are being starved of research and investment... more

34 1 reply Reply

NTTP
2 days ago

they found 1 way that worked (2 if you count reverse diffusion, which i do) and since that was the first one (2), pile-on! to the moon!

19 Reply

See all responses

More from Alberto Romero

Alberto Romero

OpenAI Researchers Have Discovered Why Language Models Hallucinate

A review of OpenAI's latest research paper

Sep 19 924 36

Alberto Romero

The Trillion-Dollar AI Bet

There are bubbles of excess and bubbles of pure betting; some bubbles are both

Oct 2 517 14

Alberto Romero

How to Live Without Your Phone

If you don’t do it for yourself, do it for your children

Oct 2 613 16

Alberto Romero

It’s Obviously the Chatbots

There’s little doubt at this point...

Oct 11 131 1

See all from Alberto Romero

Recommended from Medium

In The Generator by Thomas Smith

OpenAI Finally Admits the Real Reason it Crippled GPT-5

And what it's doing to make things right

4d ago 865 27

Max Petrusenko

The God Button: Why 40 Scientists Just Begged Us to Stop Playing Creator

One droplet could rewrite 3.5 billion years of evolution. The question isn't whether we can press this button—it's whether we should.

Oct 14 1.91K 55

Burk

Germany & Denmark Ditch Microsoft

Others will follow soon

Jun 24 3.6K 86

In Towards Deep Learning by Sumit Pandey

Why Everyone Will Want DGX Spark on Their Desk — Yes, Everyone

I just saw this picture today and was amazed, I've been waiting for this moment for a long time. (No, it's not Elon.) It's that tiny...

Oct 14 23 7

Devansh

Why Jensen Huang Loves the “AI Bubble” Stories

Answering why AI is not a Bubble and the Deeper Story at play

Oct 17 287 12

Cory Doctorow

The AI that we’ll have after AI

Cheap GPUs, unemployed engineers, and open source models.

Oct 16 1.7K 30

See more recommendations