

# The Problem with Reasoners

November 20, 2024

Aidan McLaughlin

---

o1 reasoners are the most exciting models since the original GPT-4. They prove what I predicted earlier this year in my  [AI Search: The Bitter-er Lesson](#) paper: we can get models to think for longer instead of building bigger models.

It's too bad they suck on problems you should care about.

## RL is Magic Until it Isn't

Before we built large language models, we had reinforcement learning (RL). 2015 - 2020 was the golden age of RL, and anyone remotely interested in AI stared at the ceiling wondering how you could build "AlphaZero, but general."

RL is magic: build an environment, set an objective, and watch a cute little agent ascend past human skill. When I was younger, I wanted to RL everything: Clash Royale, investment analysis, algebra; whatever interested me, I daydreamed about RL that could do it better.

Why is RL so fun? Unlike today's chatbots, it gave you a glimpse into the mind of God. AlphaZero was artistic, satisfying, and *right*; it metaphorized nature's truths with subtle strategy and stories of the cosmos with daring tactics. Today's smartest chatbots feel like talking to a hyper-educated human with 30th-percentile intelligence. AI can be beautiful, and we've forgotten that.





Stockfish 8 (Lost)

AlphaZero (black) plays Ng5, sacrificing its knight for no immediate material gain. This move was flippant but genius; it changed chess forever. After seeing it, I realized I wanted to research AI.

But RL is often useless.

In high school, I wondered how to spin up an RL agent to write philosophy essays. I got top marks, so I figured that writing well wasn't *truly* random. Could you reward an agent based on a teacher's grade? Sure, but then you'd never surpass your teacher. Sometimes, humanity has a philosophical breakthrough, but these are rare; certainly not a source of endless reward. How would one reward superhuman philosophy? What does the musing of aliens 10 × smarter than us look like? To this day, I still have no idea how to build a philosophy-class RL agent, and I'm unsure if anyone else does either.

RL is great for board games, high-frequency trading, protein folding, and sports... but not for open-ended thought without clear feedback . In formal RL, we call these problems **sparse reward environments**. While we've designed more efficient RL algorithms, there is no silver bullet to solve them.

How does o1 work? Well, nobody outside of OpenAI for sure, but my high-level guess is:

You take a language model, like GPT-4.

You gather (or generate) a large dataset of questions with known answers.

You spin up an RL environment where reasoning steps are actions, previous tokens are observations, and reward is the solution's correctness.

That's it.

Language is open-ended, but sometimes, we constrain it into environments with fixed rules. Python, for instance, either executes or doesn't. Lean, a language for formal math verification, allows you to check proof validity with a computer. Volumes of AP/SAT tests exist with answer keys. We call areas with many known answers "domains with easy verification," and they fuel o1's training pipeline.

Unlike traditional tabula rasa (from scratch) RL, o1 already understands language. It would be hard to train an RL algorithm to do advanced math, programming, or test-taking from scratch—even if there's frequent reward—because the agent must first understand language. o1 *is* magical: it allows us to do powerful RL on language games for the first time.

But, despite this impressive leap, remember that o1 uses RL, RL works best in domains with clear/frequent reward, and most domains lack clear/frequent reward.

## Praying for Transfer Learning

OpenAI admits that they trained o1 on domains with easy verification but hope reasoners generalize to all domains. Whether or not they generalize beyond their RL training is a trillion-dollar question. Right off the bat, I'll tell you my take:

o1-style reasoners do not meaningfully generalize beyond their training.

Transfer learning is the idea that models trained in one area may improve elsewhere. Earlier this year, everyone discussed multimodal transfer learning. We thought training our models on images or audio would enable better visual or language understanding, but we were wrong. [gpt-4o](#), the first large end-to-end multimodal model, scored about the same on spatial reasoning tests (like ARC bench) as its predecessors. In fact, larger *text-only* models seem to outperform multimodal [gpt-4o](#) on questions that demand a minds-eye manipulation. Today, months later, there have been zero examples of significant transfer learning success.

So does transfer learning work for reasoners? Does training on math/coding/tests allow a general inference compute unlock? DeepSeek's new [r1](#) reasoning model (trained like o1 with RL) achieves higher-than- [o1-preview](#) performance on several math/coding benchmarks. Unlike o1 models, DeepSeek exposes r1's reasoning so we can view its thought process. Let's see how it compares to a normal LLM (asked to reason in sentence-long chunks) when asked to write a Haiku. The model's **reasoning is in orange**, and its **output is in blue**. Can you tell which is the reasoner and which is the normal LLM?

### Model 1

I need to write a haiku, which requires a 5-7-5 syllable structure.

The singularity is a complex technological concept about artificial intelligence surpassing human intelligence.

I need to capture the essence of transformation and potential.

**Key themes to consider:**

### Model 2

Alright, I need to write a haiku about the singularity. Haikus are three lines with a 5-7-5 syllable structure.

The singularity usually refers to the point where AI surpasses human intelligence, so I should capture that essence.

First line: Maybe something about AI waking up or evolving.

**Key themes to consider:**

machine consciousness,  
technological explosion,  
human-AI boundary blurring.

Potential imagery: circuits,  
networks, emerging  
consciousness.

Syllable check for potential  
lines:

"Ma-chines a-wake now" (5  
syllables)

"Net-works ex-plode be-yond"  
(7 syllables)

"Con-scious-ness blooms" (5  
syllables)

Must balance technical concept  
with poetic abstraction.

Want to evoke both wonder and  
slight existential unease.

Metaphors of emergence,  
transformation are crucial.  
Contemplating technological  
sublime - moment of radical  
change.

Revisiting syllable structure  
carefully. Final syllable  
verification needed.

Circuits spark awake  
Networks bloom beyond  
knowns edge  
Thought weaves itself free

"Machine mind awakens" –  
that's 5 syllables.

Second line: I want to show the  
transformation or merging of  
human and machine. "Human  
and code entwine in dance" –  
that's 7 syllables.

Third line: The final impact or  
what happens after the  
singularity. "Future echoes, vast  
and strange" – that's 5  
syllables.

Let me check the syllables  
again to make sure:

Machine mind awakens (5)

Human and code entwine in  
dance (7)

Future echoes, vast and  
strange (5)

Perfect, it fits the haiku  
structure and conveys the  
theme of the singularity.

Machine mind awakens,  
Human and code entwine in  
dance,  
Future echoes, vast.

I certainly cannot tell which model is the reasoner and which is a  
normal LLM. **Model 1** is [claude-3.5-haiku](#), and **Model 2** is  
DeepSeek's new [r1](#). (Note that r1 also uses an incorrect number of  
syllables for every line.) [r1](#) reasoning process suggests that, for most  
non-coding/math problems these models reason no differently than

asking a standard model to think step-by-step. The skills learned in narrow domains *do not* transfer to all domains.

r1 even abandons the prose it uses on math problems when tackling problems outside its RL training. The *style* doesn't even transfer, let alone performance:

**Aidan McLau**  
@aidan\_mclau · [Follow](#)

as [@TheXeophon](#) pointed out, r1 excitedly hops into lowercase when doing math and writes normally when thinking about basically any other task

this is so fascinating to me

Last edited 10:55 AM · Nov 23, 2024

---

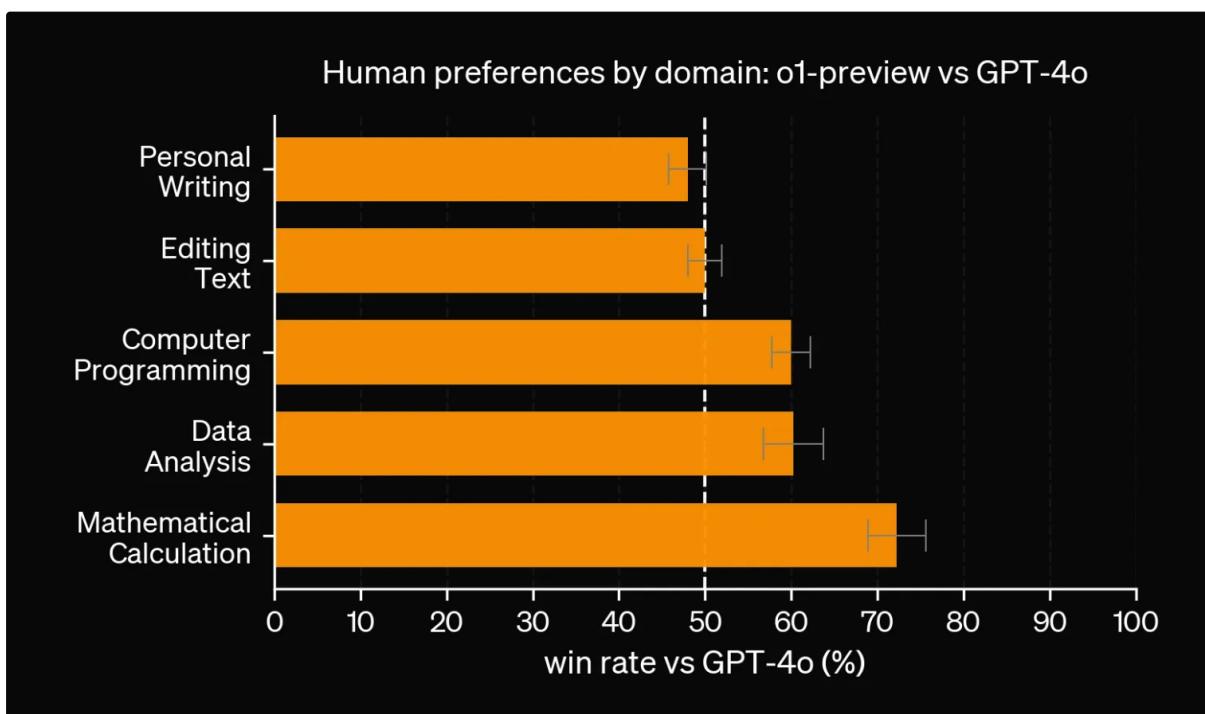
95    [Reply](#)    [Copy link](#)

[Read 8 replies](#)

We can only view summaries of o1's reasoning traces but observe something similar. On math/coding problems, users have spotted o1 hopping into Japanese, talking about random fruit, or going off on tangents. On out-of-distribution tasks, its reasoning is just summary of OpenAI's guidelines and an overview of its response.

A straightforward way to check how reasoners perform on domains without easy verification is benchmarks. On math/coding, OpenAI's o1

models do exceptionally. On everything else, the answer is less clear.



OpenAI shows that `o1-preview` does worse on writing than `gpt-4o`, a  $6 \times$  cheaper model.

Results that jump out:

`o1-preview` does worse on personal writing than `gpt-4o` and no better on editing text, despite costing  $6 \times$  more.

OpenAI didn't release scores for `o1-mini`, which suggests they may be worse than `o1-preview`. `o1-mini` also costs more than `gpt-4o`.

On eqbench (which tests emotional understanding), `o1-preview` performs as well as `gemma-27b`.

On eqbench, `o1-mini` performs as well as `gpt-3.5-turbo`. No you didn't misread that: it performs as well as `gpt-3.5-turbo`.

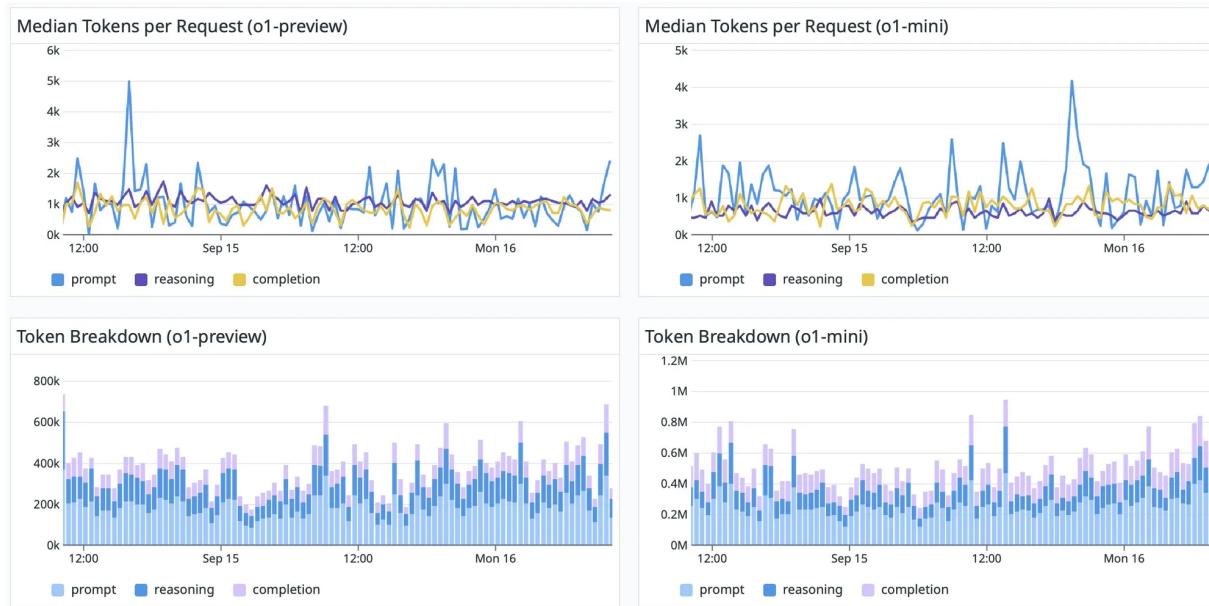
Nathan Lambert notes that this isn't surprising to post-training experts.

When asked what happened when he trains models with RL, he remarked:

So like at this point of training, normal evaluations have totally tanked. The model is like not as good at normal things.

RL-based reasoners also don't generalize to longer thought chains. In other domains, you can scale inference compute indefinitely. If you want to analyze a chess position, you can set Stockfish to think for 10

seconds, minutes, months, or eons. It doesn't matter. But the longest we observe o1-mini/preview thinking for is around 10k tokens (or a few hours of human writing). This is impressive, but to scale to superintelligence, we want AI to think for centuries of human time.



o1-mini/preview token data from OpenRouter. Note that the reasoning tokens for both models hang around 1k, which is what a normal LLM would produce when asked to think step-by-step.

When I talked to OpenAI's reasoning team about this, they agreed it was an issue, but claimed that more RL would fix it. But, as we've seen earlier, scaling RL on a fixed model size seems to eat away at other competencies! The cost of training o3 to think for a million tokens may be a model that *only* does math.

Scaling RL to unlock longer thought chains defeats the whole purpose of scaling inference compute. Like a video game potion that endows you with a temporary buff, inference compute allows you to snatch tomorrow's AI capabilities by paying more today. It expedites the future and avoids the wait. If OpenAI needs more RL to unlock higher inference ceilings, you're back to waiting for future models.

But don't let me convince you that OpenAI's o1 models suck.

o1-preview (tied with [claude-3.5-sonnet](#)) is my single favorite model. I've paid for an absurd amount of [o1-preview](#) tokens, and it was money well-spent. In my experience, o1-preview *is* great for out-of-distribution tasks. The likely explanation, however, isn't that reasoning enabled creativity, but rather that o1-preview is the single largest model

OpenAI's released since the original [gpt-4-0314](#) (interestingly, both models behave and are priced similarly). At \$15 per 1M input tokens and \$60 per 1M output tokens, it's also the most expensive modern model. But for my use, I've switched to [claude-3.5-sonnet](#), which feels more generally capable.

## Why You Should Care

The evidence suggests that RL-based reasoners struggle to perform on domains outside their training. If you agree, you still might wonder:

We can scale both model size and reasoning, right?

Do I care that RL *only* works well on domains with easy verification?

## Deep Learning May Actually Hit A Wall

The industry is having a tough time shipping larger-than-GPT-4 models. In November 2023, I thought we'd get GPT-4.5 by December 2023. A year later, in November 2024, I'm unsure if we'll ever see GPT-4.5.

I honestly didn't expect this.

If you follow me on Twitter, you'll know I was incredibly bullish on model scaling. The performance trends looked clear, and (absent a data wall) it seemed easy to add more parameters/data and see predictably better results. But, since [gpt-4-0314](#), it's unclear if we've seen a *single model larger* ([claude-3-opus](#) may be an exception).

That's wild! Sam Altman, Dario Amodei, and Demis Hassabis talked for the last 18 months about building larger models, and despite a year of miracles, a larger model is the one thing we haven't seen.

Why? I suspect a few reasons:

Post-training is increasingly tricky on models that have developed worldviews.

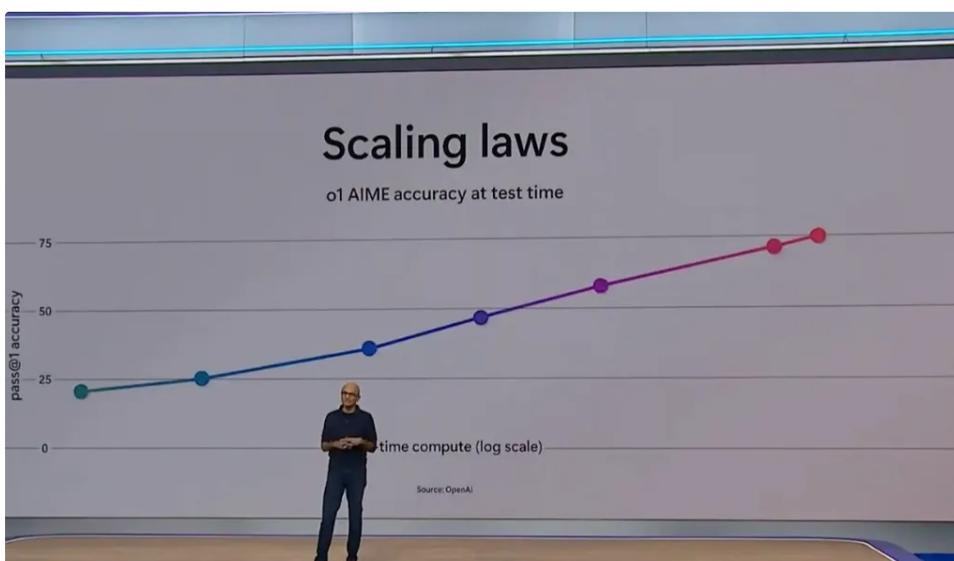
For every model you see, there's a graveyard of failures. Bigger models are slower and more expensive to train; thus, fewer shots and no shots on target exist.

Big models are expensive to deploy compared to their smaller peers, making it challenging to justify when benchmark performance is similar.

Even lab rhetoric has shifted.

Instead of "there's no end in sight" for building bigger models, we now hear, "We have many scaling directions left."

Instead of "We have a high degree of scientific certainty that GPT-5 will be better than GPT-4," discussion on GPT-5 has basically stopped.



Satya talks about "a new scaling direction" in case pre-training scaling slows down.

The world seems to have abandoned building larger models, and this breaks my heart. Massive language models are the Lunar missions of the 21st century. Like stepping on another celestial body, they open the heavens a bit wider to humanity. Models like [gpt-4-0314](#) and [claude-3-opus](#) approach *things*, not just impressive tools. Scientists turned sand and electricity into borderline sentient peers. But, like the Lunar mission, I fear the investment return is too unclear or the science too unreliable to fuel their growth.

I hope we don't enter a big model winter, but things don't look good. Everyone outside OpenAI has bent their rhetoric, research talent, and

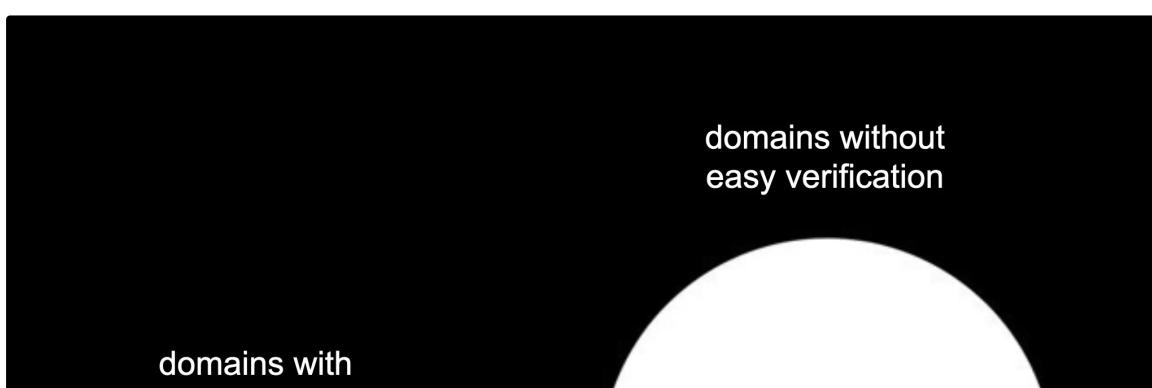
thought on duplicating o1. All of OpenAI's efforts have turned to training o2. Sure, o1-successors may crack Millennium Prize Problems, but without scaling model size, that may be all.

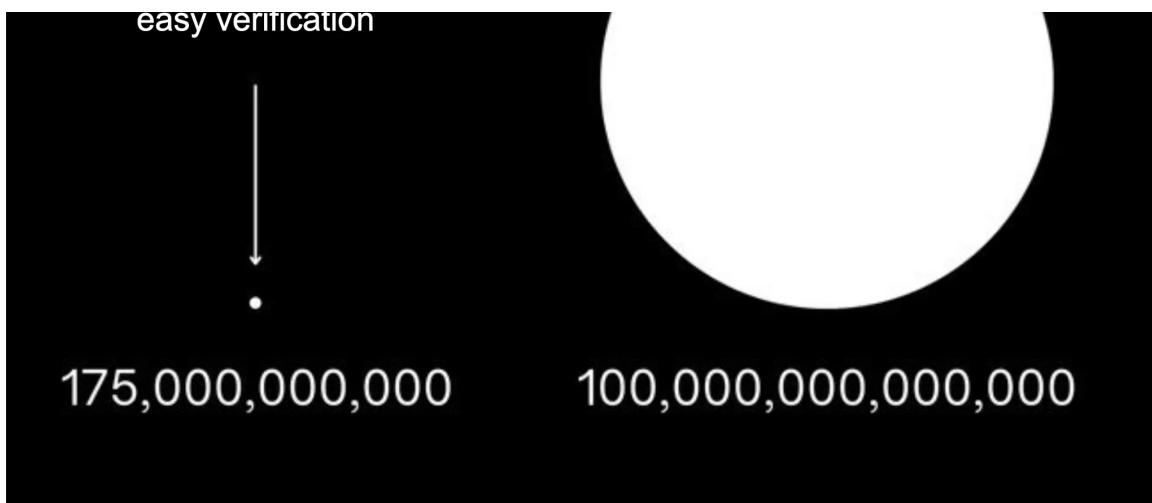
## Domains *Without* Easy Verification

Throughout this essay, I've doomsayed o1-like reasoners because they're locked into domains with easy verification. You won't see inference performance scale if you can't gather near-unlimited practice examples for o1.

If you're not a mathematician or Leetcode wizard, you probably care a lot more about domains without easy verification. If we only build better reasoners, and not larger models, AI will forever be stuck at GPT-4-level for:

- Giving corporate strategy
- Soothing a hurt friend
- Advising governments
- Writing a hit philosophy paper
- Startup investment
- Managing a team
- Writing brilliant poetry
- Giving career advice
- Understanding social trends
- Giving feedback on an essay





Please consult my rigorous graph.

Even for coding and math, it's unclear if reasoners will be that useful. Great programmers don't just write bug-free code that executes fast; they launch companies and exercise aesthetic and conceptual taste. Famous mathematicians don't just churn out proofs; they wonder about the world and invent new mathematics that gives them explanatory power.

You may be tempted to think that inference-time compute scaling is a hoax and that AI's stuck, but it's worth noting that humans think for extended periods on open-ended questions. Einstein stared at ceilings and imagined racing light beams until he spawned relativity. Essayists mull for months until an eloquent idea pops out. Students iteratively refine presentations until they end with something better than they started. Poets spend years obsessing over a handful of lines.

Do you think Einstein was trained on countless decade-long reasoning chains? When you write an essay, do you transfer skills from mathematics to make it better? When you think deeply about responding to a friend, do you pull on your extensive practice minimizing program execution time?

## Conclusion

While reading this, you may suspect I'm an AI skeptic.

I'm not.

I expect transformative AI to come remarkably soon. I hope labs iron out the wrinkles in scaling model size. But if we do end up scaling model size to address these changes, what was the point of inference compute scaling again?

Remember, inference scaling endows today's models with tomorrow's capabilities. It allows you to skip the wait. If you want faster AI progress, you want inference to be a 1:1 replacement for training.

o1 is not the inference-time compute unlock we deserve.

If the entire AI industry moves toward reasoners, our future might be more boring than I thought.

(Thanks to Emil, Neel, and Nicole for feedback on drafts.)