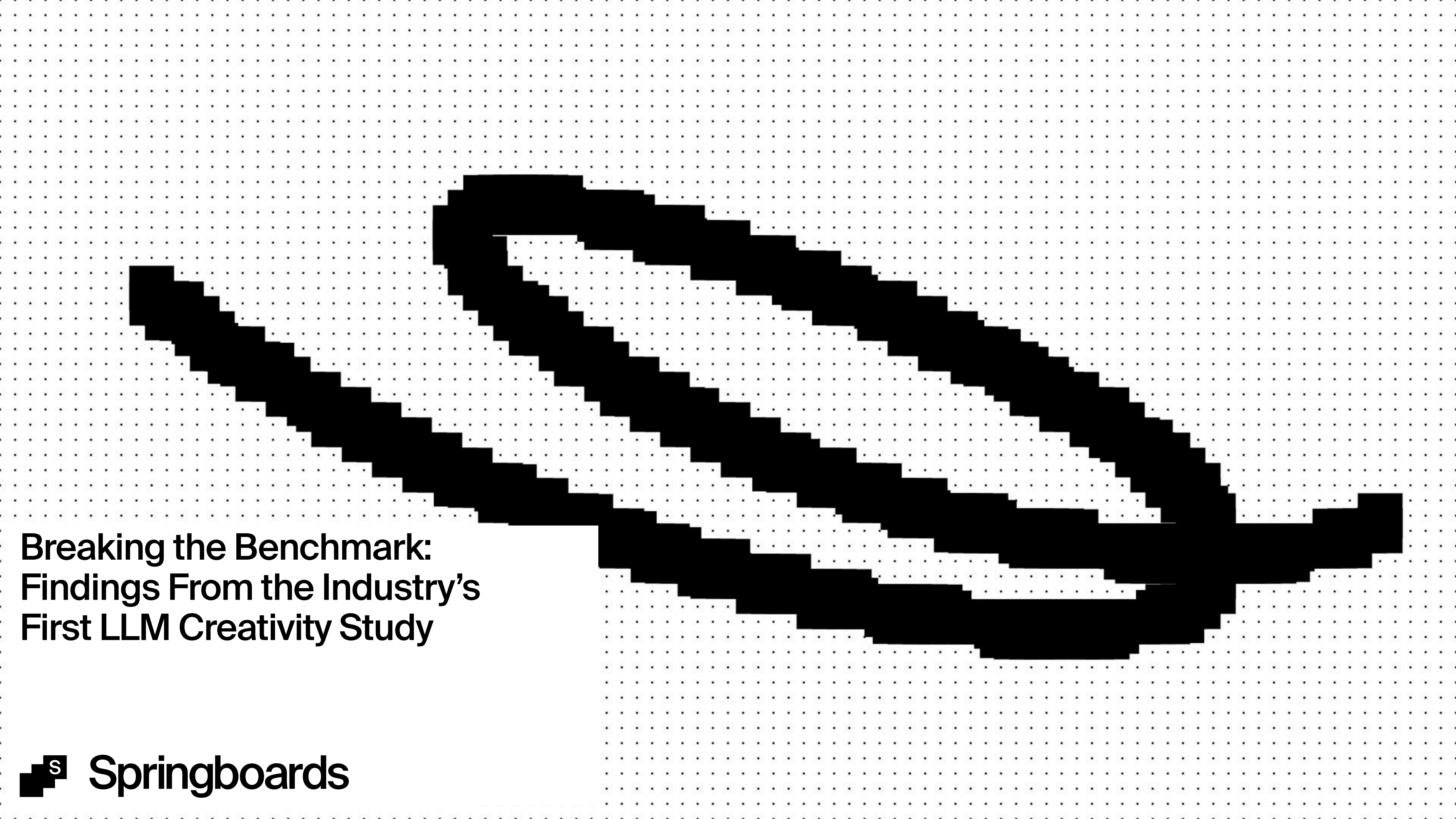




**Breaking the Benchmark:
Findings From the Industry's
First LLM Creativity Study**

↓
SPRINGBOARDS IS

**A creative tool to inspire,
not give you the answers.**



↓
AND YES, WE ARE AN AI TOOL..

**but a self
loathing one**





SAM ALTMAN

“AI will handle 95% of what marketers use agencies, strategists, marketing and creative professionals for today.”

March 2024





MARK ZUCKERBERG

“We’re going to get to a point where you’re a business, you come to us, you tell us what your objective is, you connect to your bank account, you don’t need any creative, you don’t need any targeting demographic, you don’t need any measurement, except to be able to read the results that we spit out”

May 2025

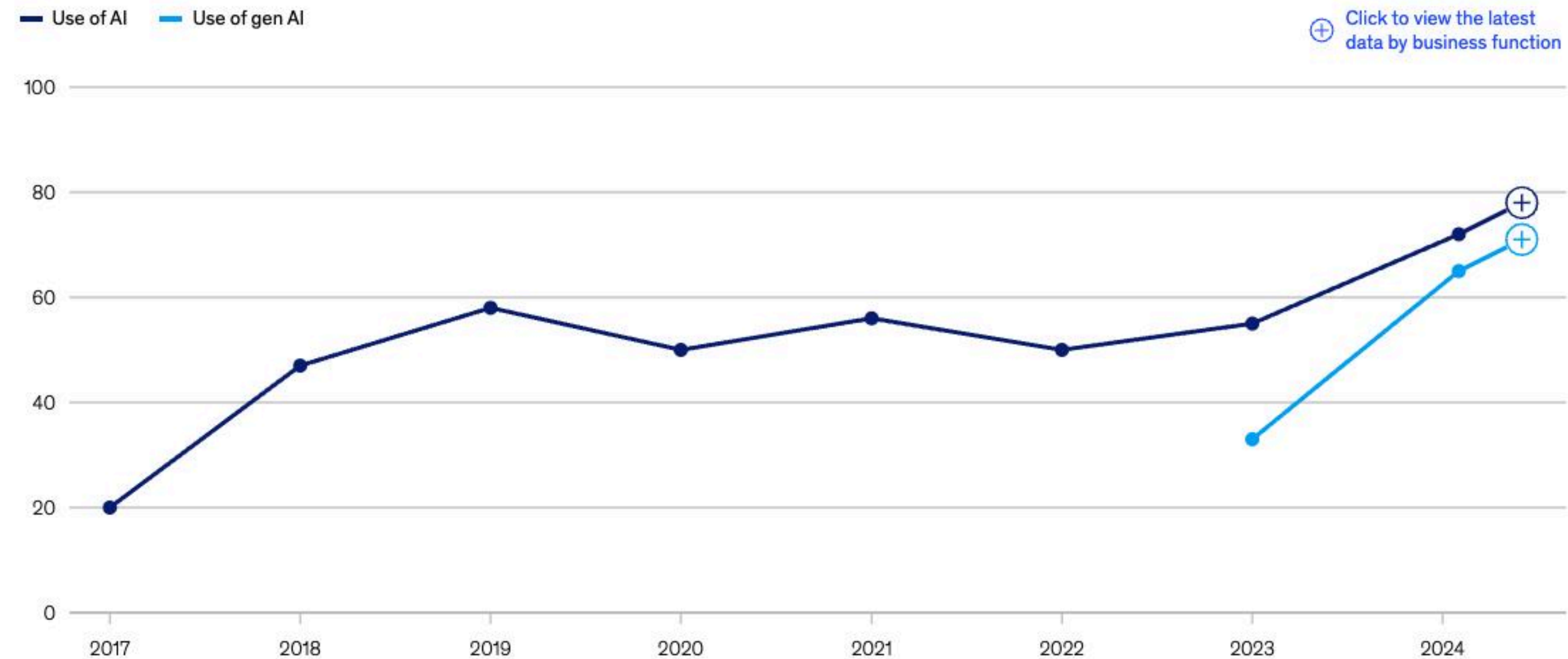


↓
THE PROBLEM

AI adoption has been rampant.

Organizations' use of AI has accelerated markedly in the past year, after years of little meaningful change.

Organizations that use AI in at least 1 business function,¹ % of respondents



¹In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.

Source: McKinsey Global Surveys on the state of AI <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>



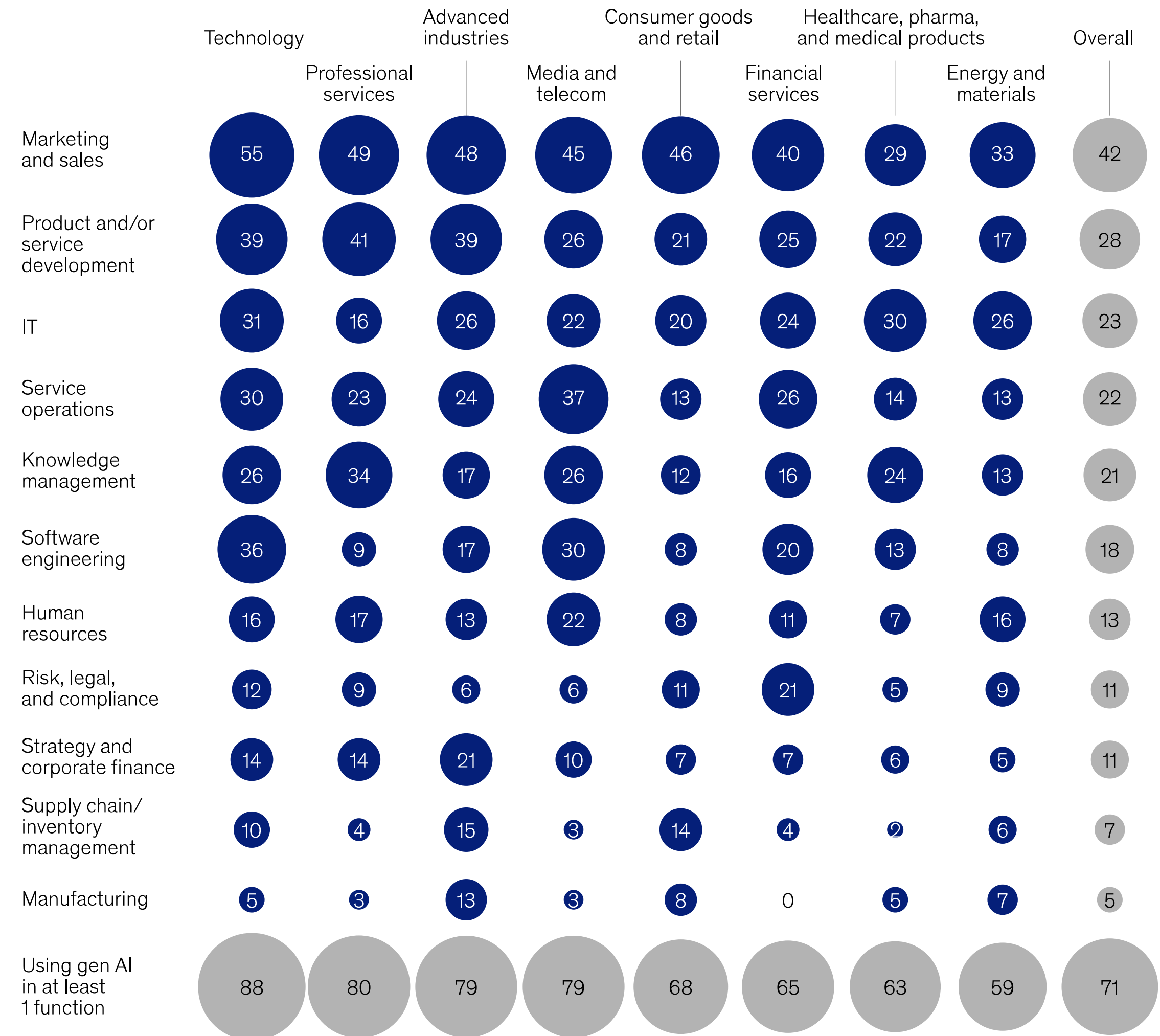
↓
THE PROBLEM

And highest
amongst marketing
teams



Organizations across industries have begun to use gen AI in marketing and sales, though other uses vary by industry.

Business functions in which respondents' organizations are regularly using gen AI, by industry,¹
% of respondents



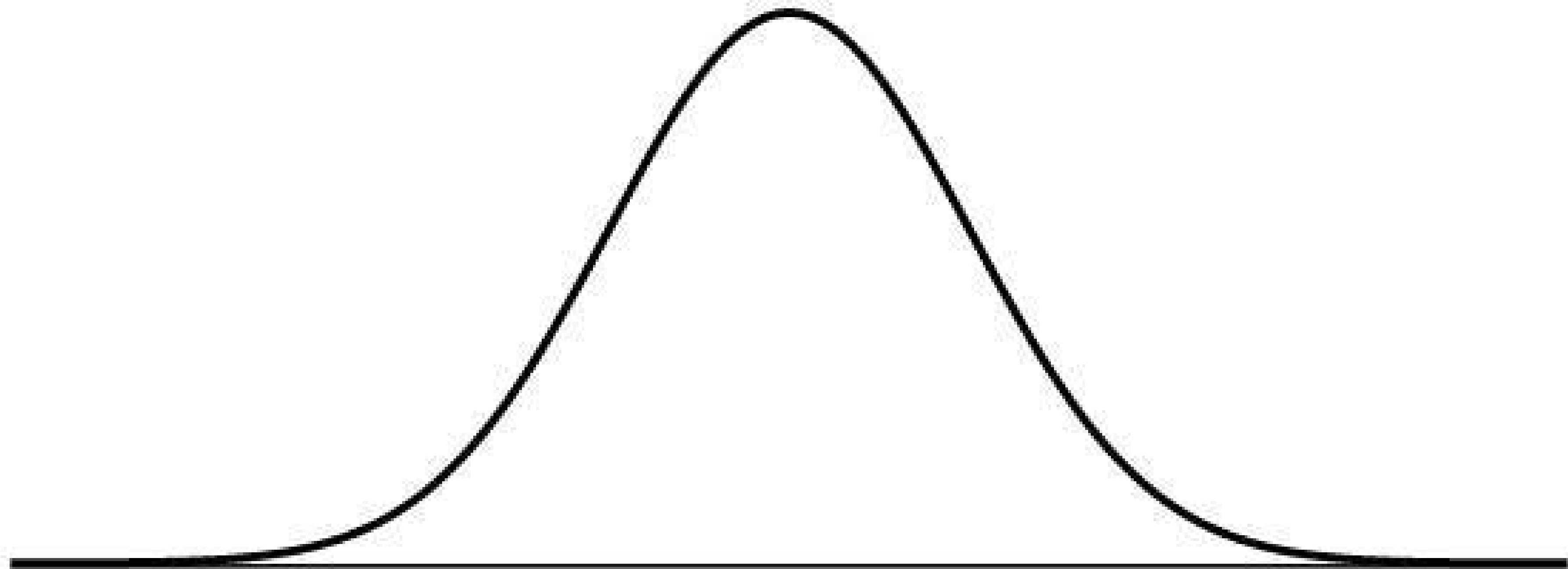
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

The background is a dark, monochromatic illustration. It depicts several hands holding smartphones. One phone in the upper left shows a social media profile with two avatars. Another phone in the lower left shows a video call with a woman's face. A third phone on the right shows a news article with a large 'A' and a picture of a mountain. The overall style is sketchy and artistic, with a focus on digital connectivity and surveillance.

A terrifying thought

↓
THE PROBLEM

**Unknown to many...
they are funneling us all into the same place.**





BECAUSE IT THREATENS TO FLATTEN CULTURE

We are seeing it happen in music

[Home](#)[Music](#)[The Latest](#)[News](#)[Live Music](#)[Programs](#)[Videos](#)[Touring](#)[Merch](#)[Hottest 100](#)[PLAY LIVE RADIO](#)

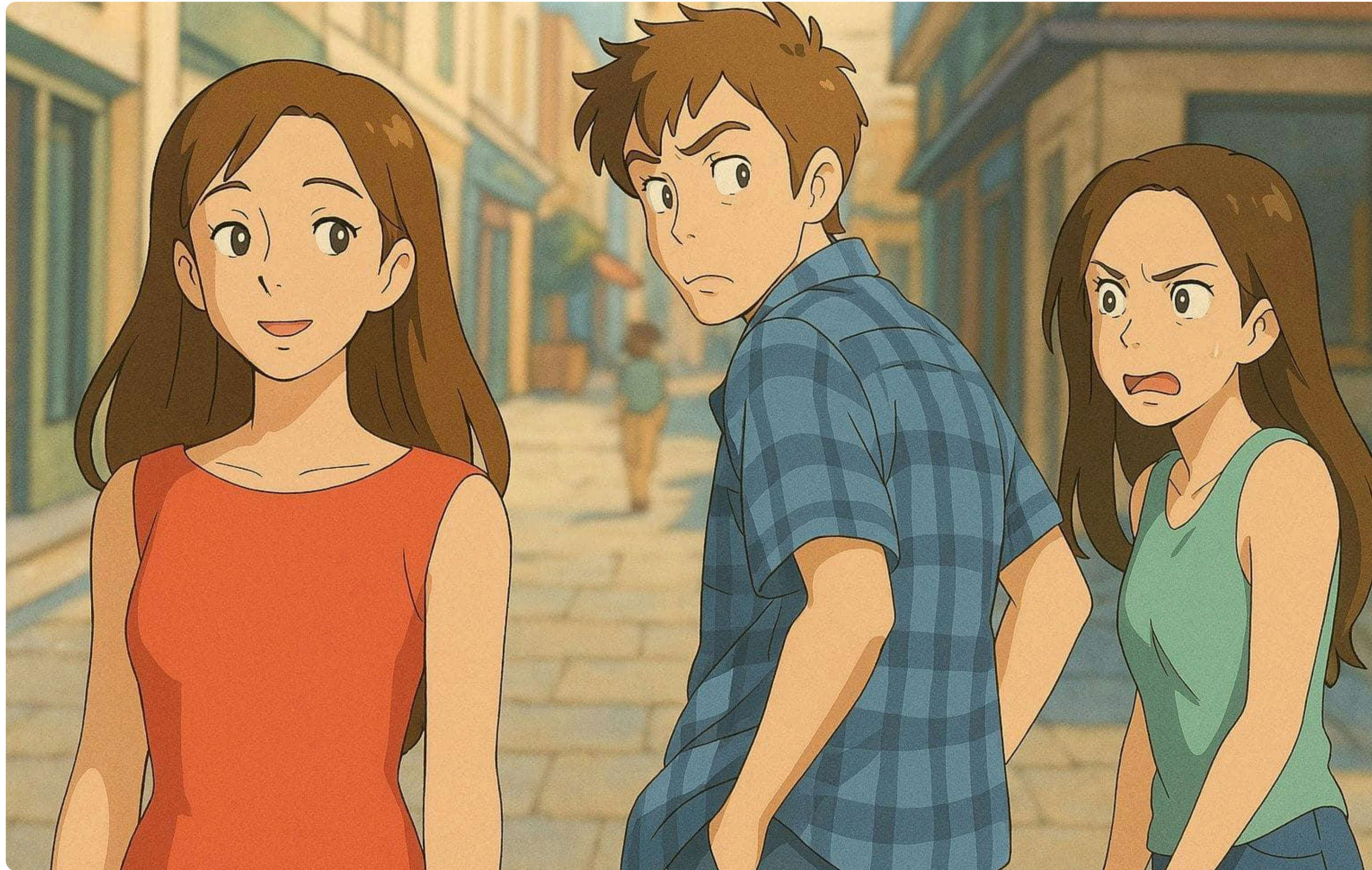
Music streamers are serving you AI-generated and 'ghost' music. Here's how it hurts real artists

By [Jared Richards](#)
Posted 28 Aug 2025



↓
BECAUSE IT THREATENS TO FLATTEN CULTURE

We are seeing it happen in culture



↓
BECAUSE IT THREATENS TO FLATTEN CULTURE

We are seeing it happen in politics

ChatGPT triggers surge in MPs using AI-written speeches

Use of common artificial intelligence phrases such as 'I rise today' has jumped since 2022

 **James Titcomb**
Technology Editor

Related Topics
ChatGPT, Artificial Intelligence,
House of Commons

11 September 2025 2:29pm BST

  44

 Gift this article free

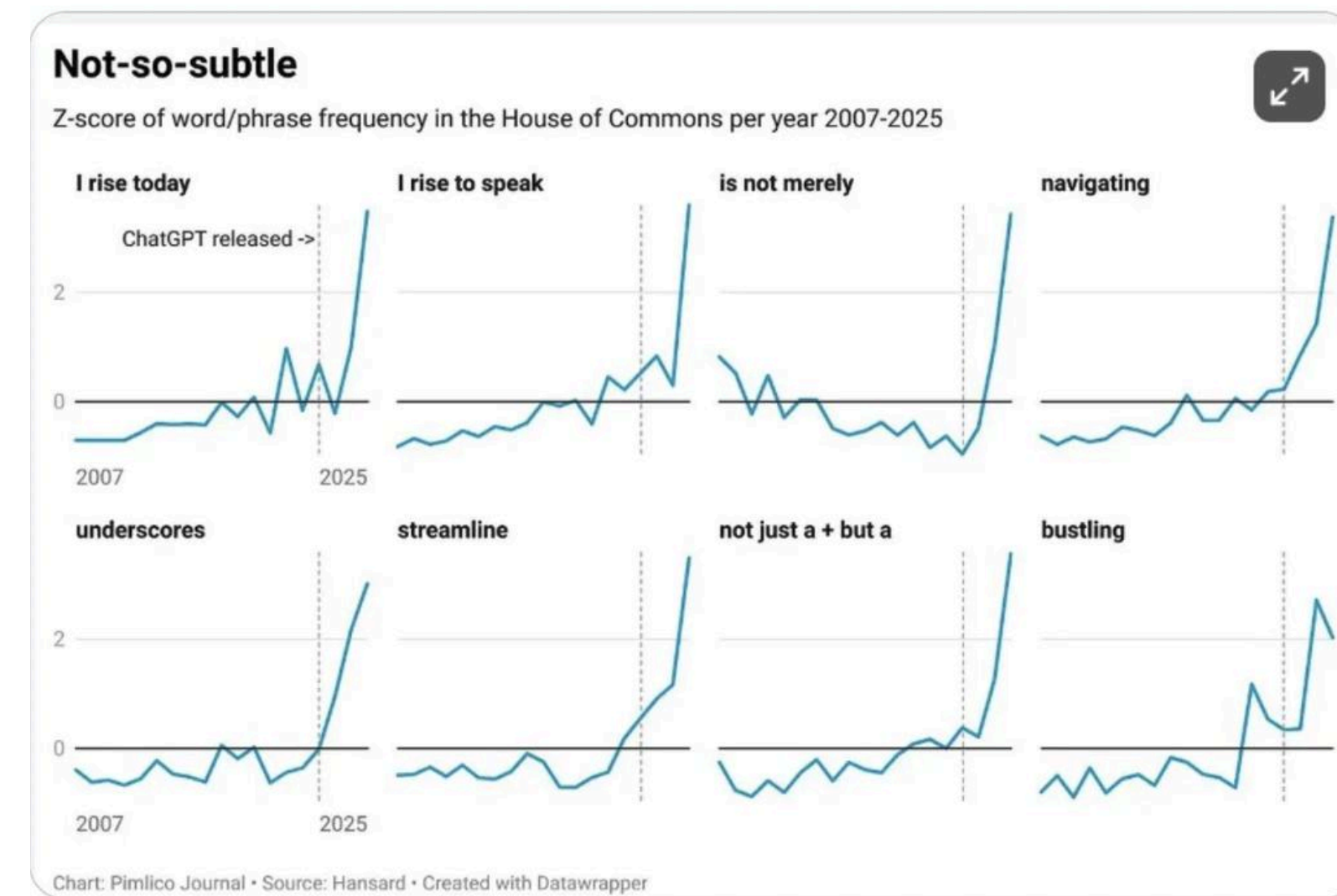


telegraph.co.uk/business/2025/09/11/chatgpt-triggers-surge-in-mps-using-ai-written-speeches/

 **François Valentin** 
@Valen10Francois

The surge in AI-written speeches in Britain's House of Commons visualised:



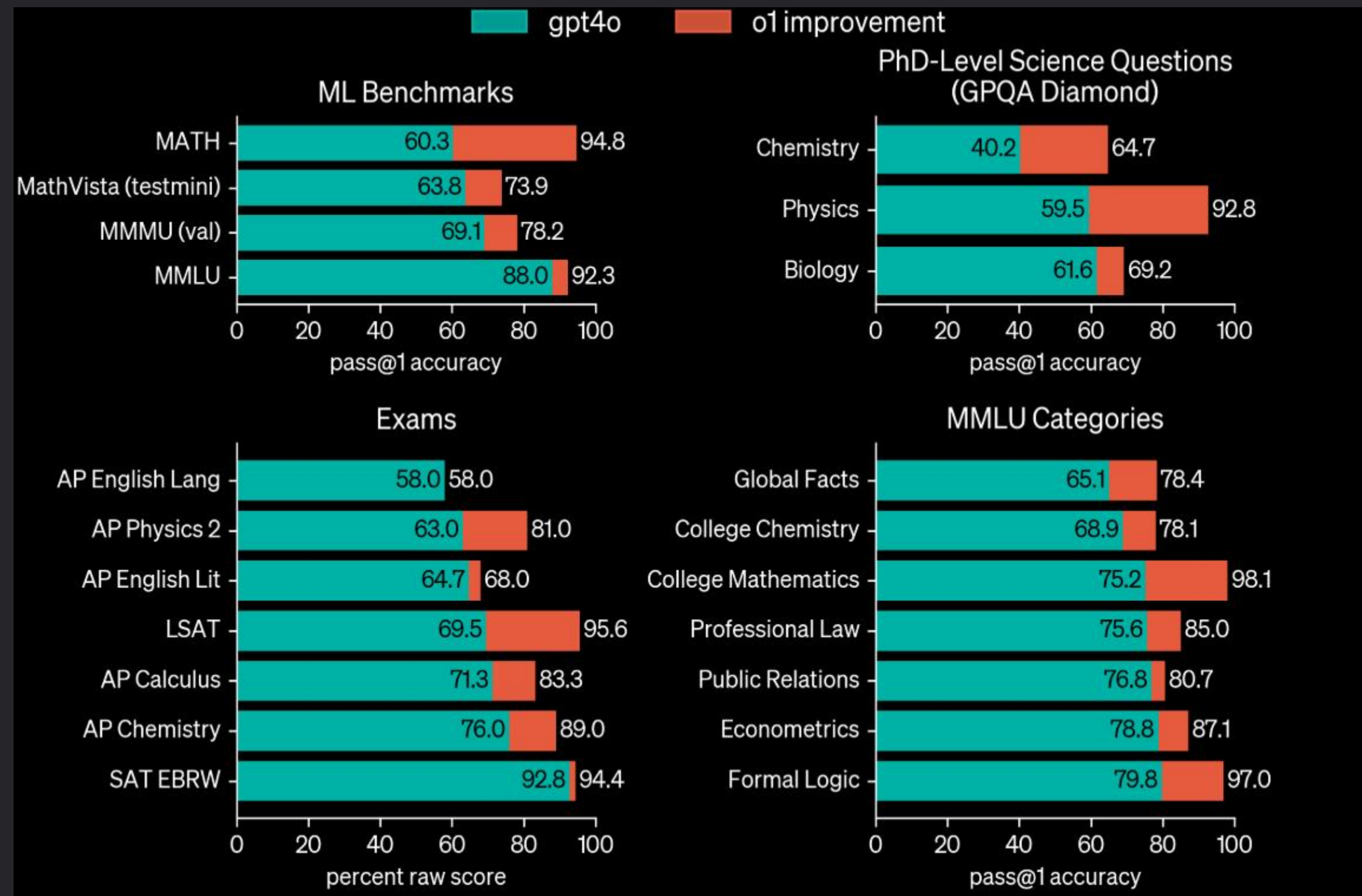
**So how do we
break the
machines?**





OPENAI EVALS

The first step is to break the benchmarks



openai.com/index/learning-to-reason-with-llms



↓
ANTHROPIC EVALS

	Claude 3.7 Sonnet <i>64K extended thinking</i>	Claude 3.7 Sonnet <i>No extended thinking</i>	Claude 3.5 Sonnet (new)	OpenAI o1 ¹	OpenAI o3-mini ¹ <i>High</i>	DeepSeek R1 <i>32K extended thinking</i>	Grok 3 Beta <i>Extended thinking</i>
Graduate-level reasoning <i>GPQA Diamond³</i>	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%	79.7%	71.5%	80.2% / 84.6%
Agentic coding <i>SWE-bench Verified²</i>	—	62.3% / 70.3%	49.0%	48.9%	49.3%	49.2%	—
Agentic tool use <i>TAU-bench</i>	—	Retail 81.2%	Retail 71.5%	Retail 73.5%	—	—	—
	—	Airline 58.4%	Airline 48.8%	Airline 54.2%	—	—	—
Multilingual Q&A <i>MMMLU</i>	86.1%	83.2%	82.1%	87.7%	79.5%	—	—
Visual reasoning <i>MMMU (validation)</i>	75%	71.8%	70.4%	78.2 %	—	—	76.0% / 78.0%
Instruction-following <i>IFEval</i>	93.2%	90.8%	90.2%	—	—	83.3%	—
Math problem-solving <i>MATH 500</i>	96.2%	82.2%	78.0%	96.4%	97.9%	97.3%	—
High school math competition <i>AIIME 2024³</i>	61.3% / 80.0%	23.3%	16.0%	79.2% / 83.3%	87.3%	79.8%	83.9% / 93.3%

<https://www.anthropic.com/news/claude-3-7-sonnet>



The creative and
advertising industry is
left in the dark.



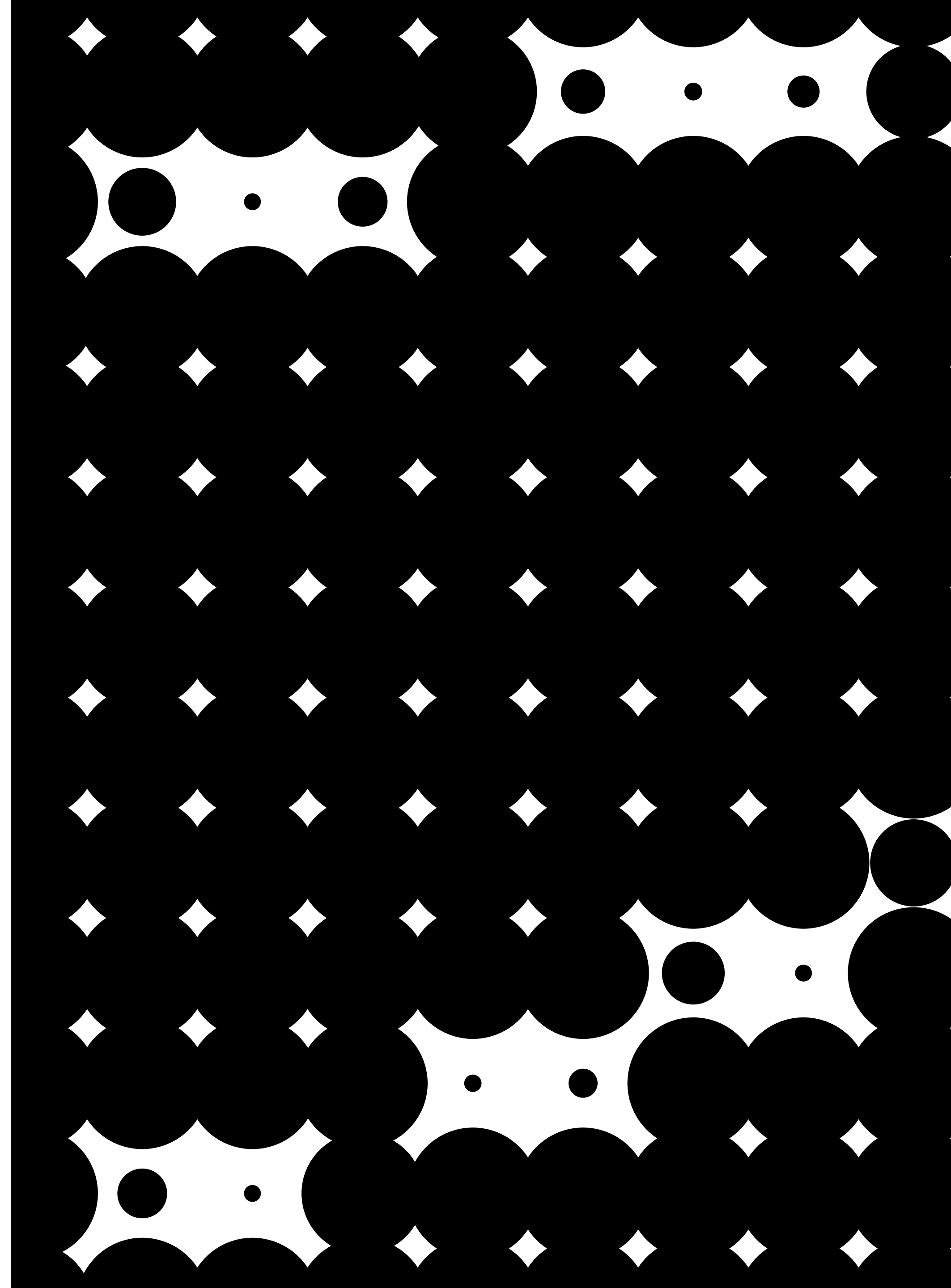
OUR VISION FOR THIS PROJECT

**A collective, industry
benchmark to help
push creativity forward**



↓
PART ONE

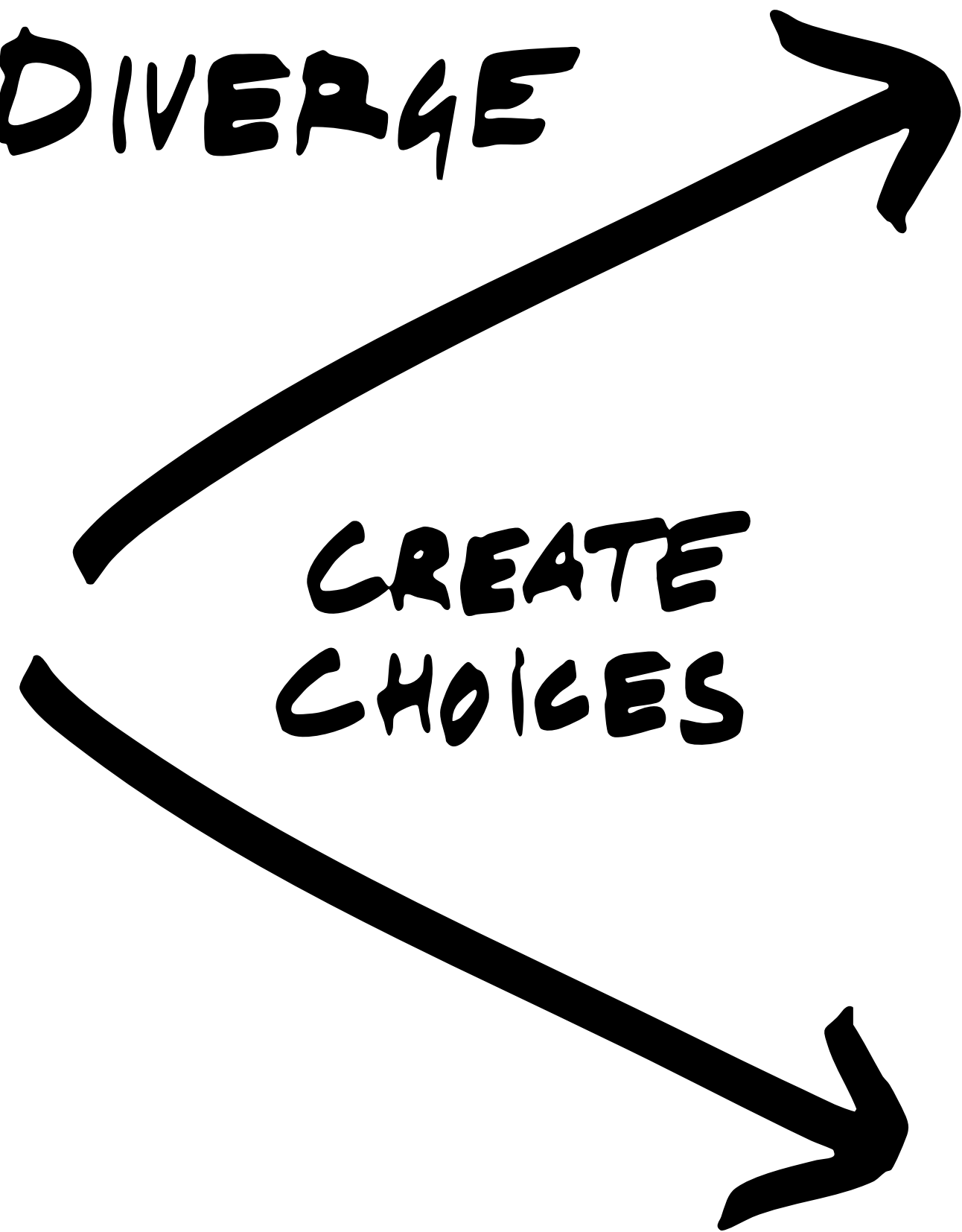
Breaking the Benchmarks



**How do you
even measure
creativity?**

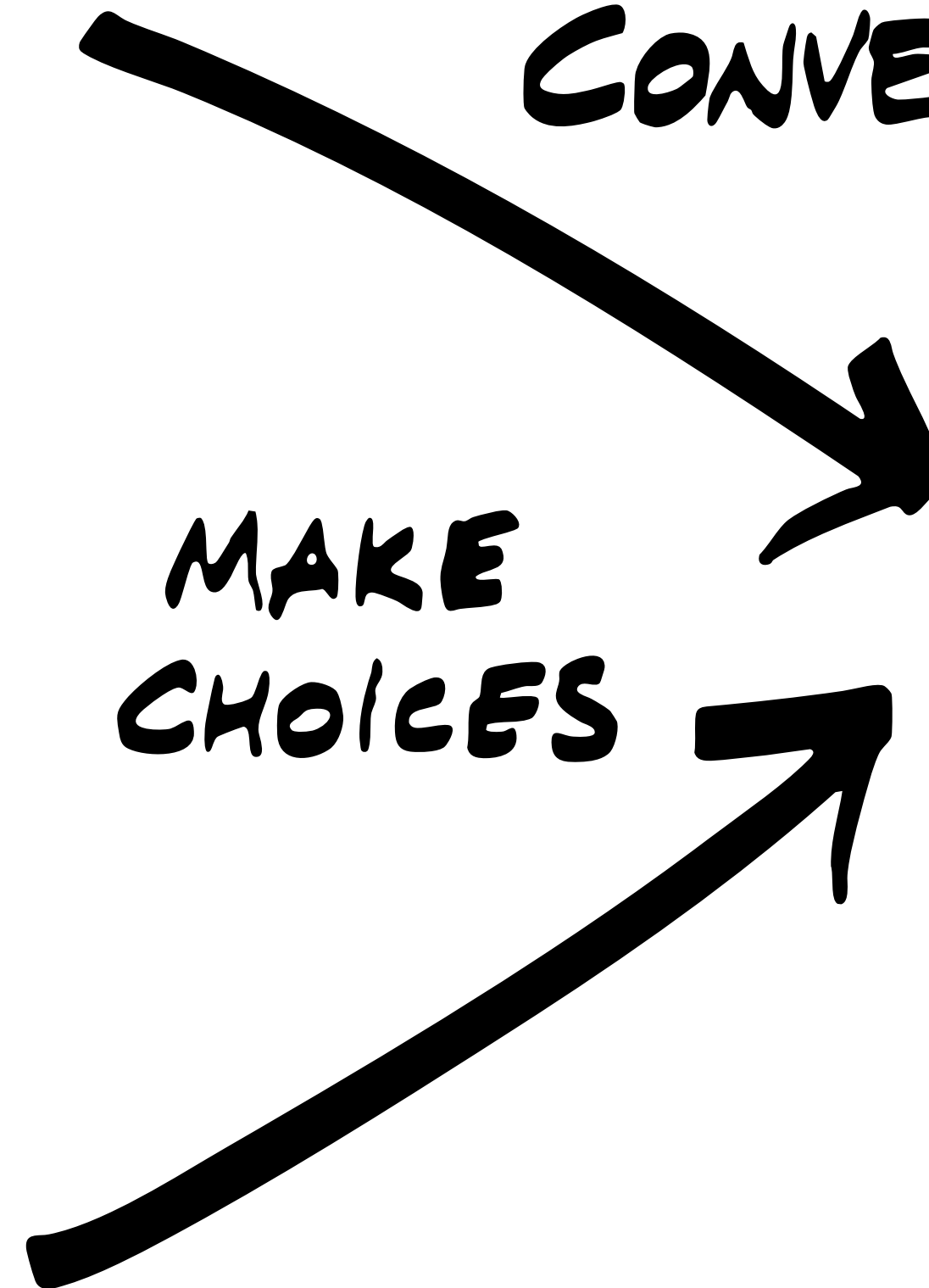


DIVERGE



**CREATE
CHOICES**

CONVERGE



**MAKE
CHOICES**

Creativity Benchmark

A collective industry benchmark for creativity.

Most AI benchmarks test for maths, science, and law, but none measure creativity. So we're building the first benchmark that reflects how our world works. Created by and for strategists, creatives, and marketers and backed by leading industry bodies. This is your chance to shape how creativity gets measured.

New here? Get started

Sign back in

Backed by



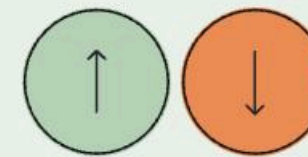
How it works

Vote on which response is better

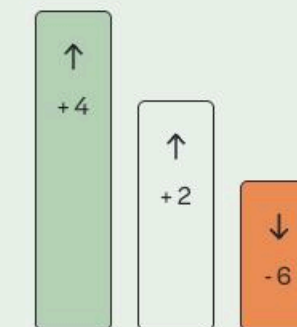
↓ Model X



↑ Model Y



Model performance is measured



View your model preferences

↓
TINDER FOR IDEAS

An industry initiative created by [Springboards.ai](#)

Creativity Benchmark

What

This initiative is a collective, industry-sourced effort to help advance creativity, because what gets measured gets better.

How

1. Vote on which response is better

2. Model performance is measured

3. View your model preferences

Sign in

Create account

Backed by

An industry initiative created by [Springboards.ai](#)

Creativity Benchmark

User Details

1 / 2

Age (select range from list)

Select

Gender (select from list)

Select

What job role do you most closely align with?

Select

Country (select from list)

Select

Continue

I already have an account

CB

You're reviewing a creative idea generated by two ML models for a specific brand.
Tap to pick the stronger one.

Brief:

Creative Insight for Nike

Response X

Nike silently buys thousands of empty storefronts

Response Y

Nike quietly buys thousands of empty storefronts

They're too similar

Skip this one

Save & Exit

2 left to unlock a new badge

An industry initiative created by [Springboards.ai](#)

Brief:

Creative Insight for Nike

↑ Gemini 2.5 Pro

Nike silently buys thousands of empty storefronts

↓ GPT 4.1

Nike quietly buys thousands of empty storefronts

Continue

Save & Exit

1 left to unlock a new badge

An industry initiative created by [Springboards.ai](#)

Creativity Benchmark

Welcome, Ronan!

200 Evals

Beyond Measure

Results

My #1

DeepSeek R1

↑ DeepSeek R1

↑ Gemini 2.5 Pro

↓ GPT 4.1

↓ LLaMA4

Top & bottom model preferences after 203 evaluations

Share:

Want to refine your results & level up?

Continue Test

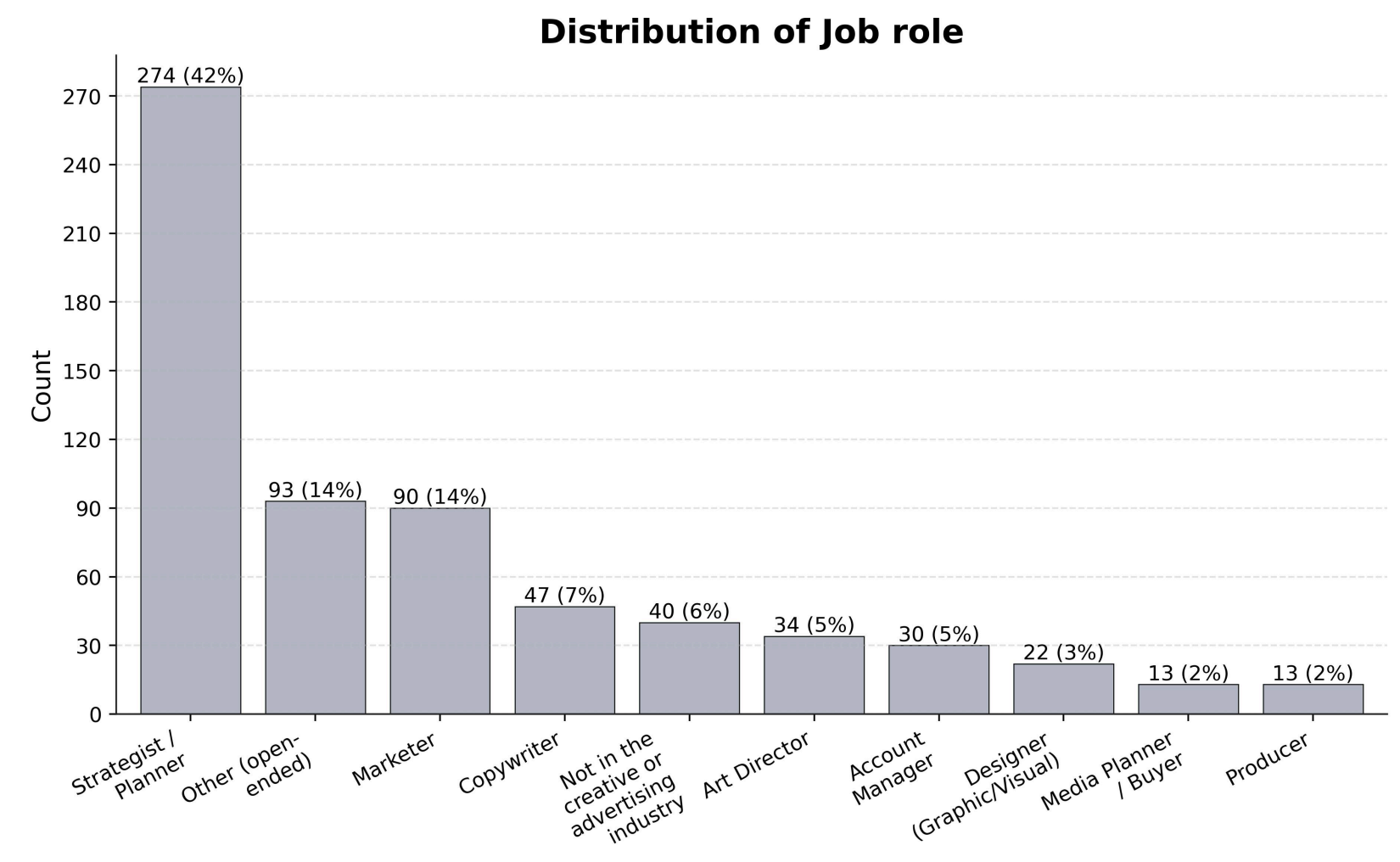
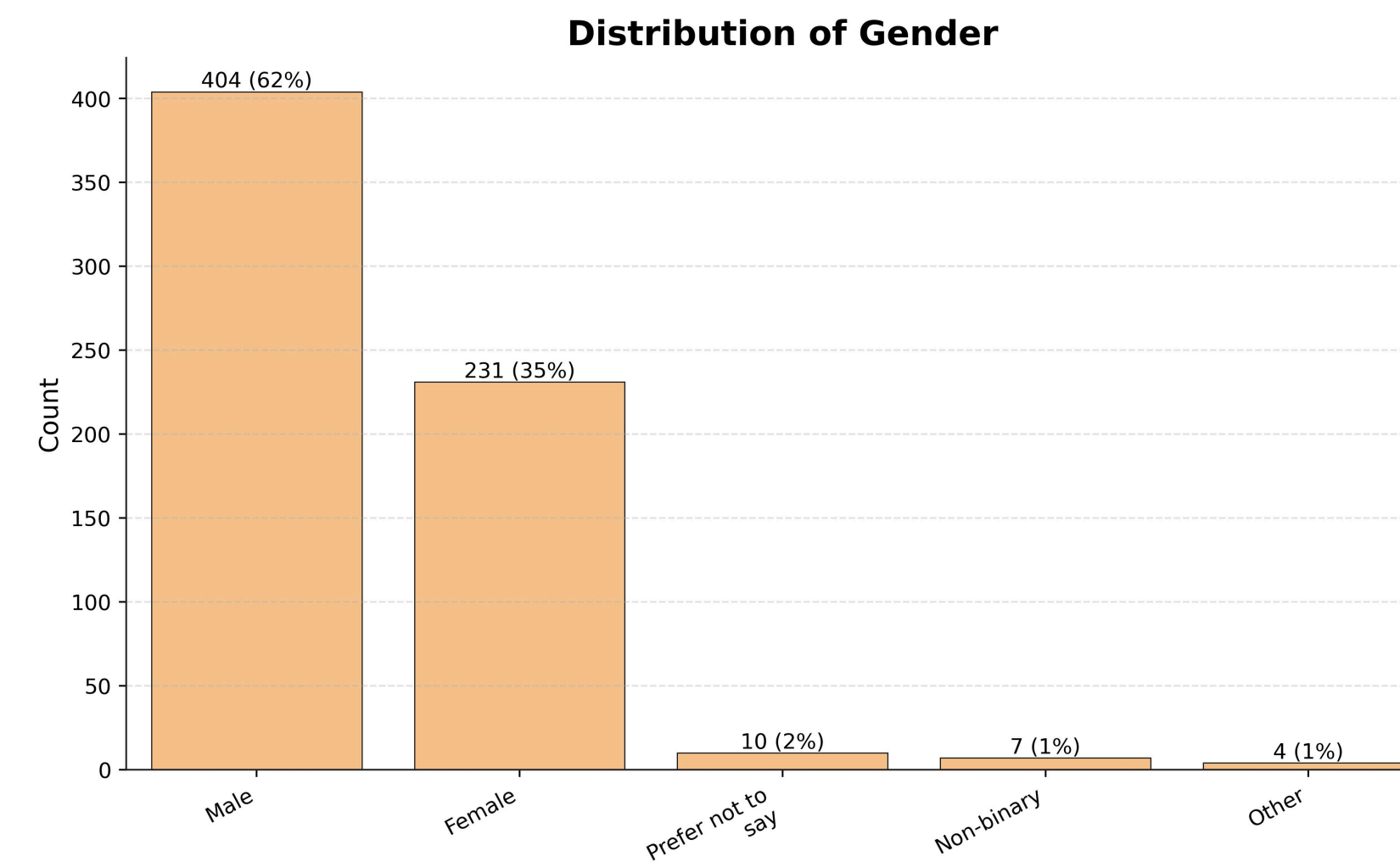
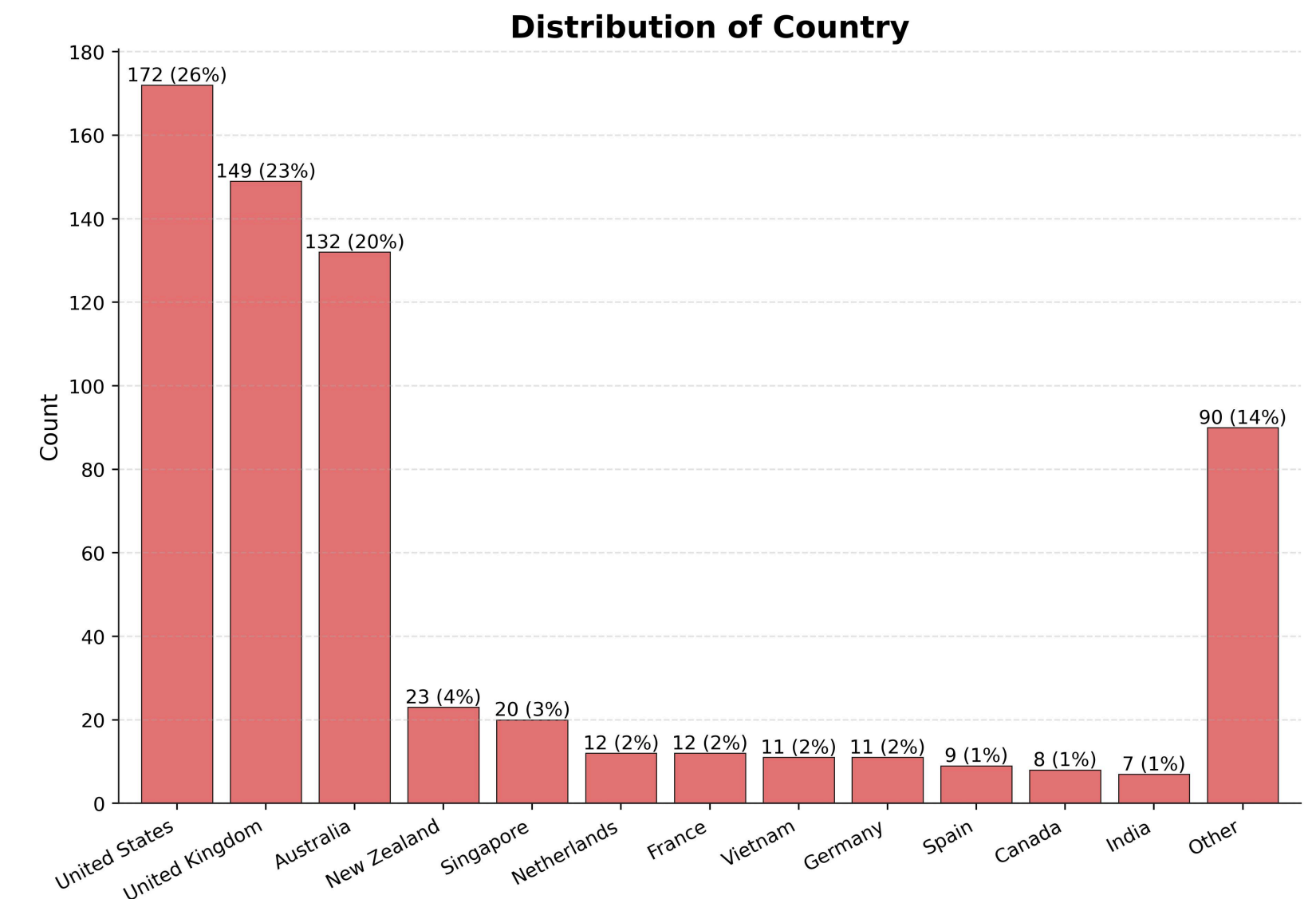
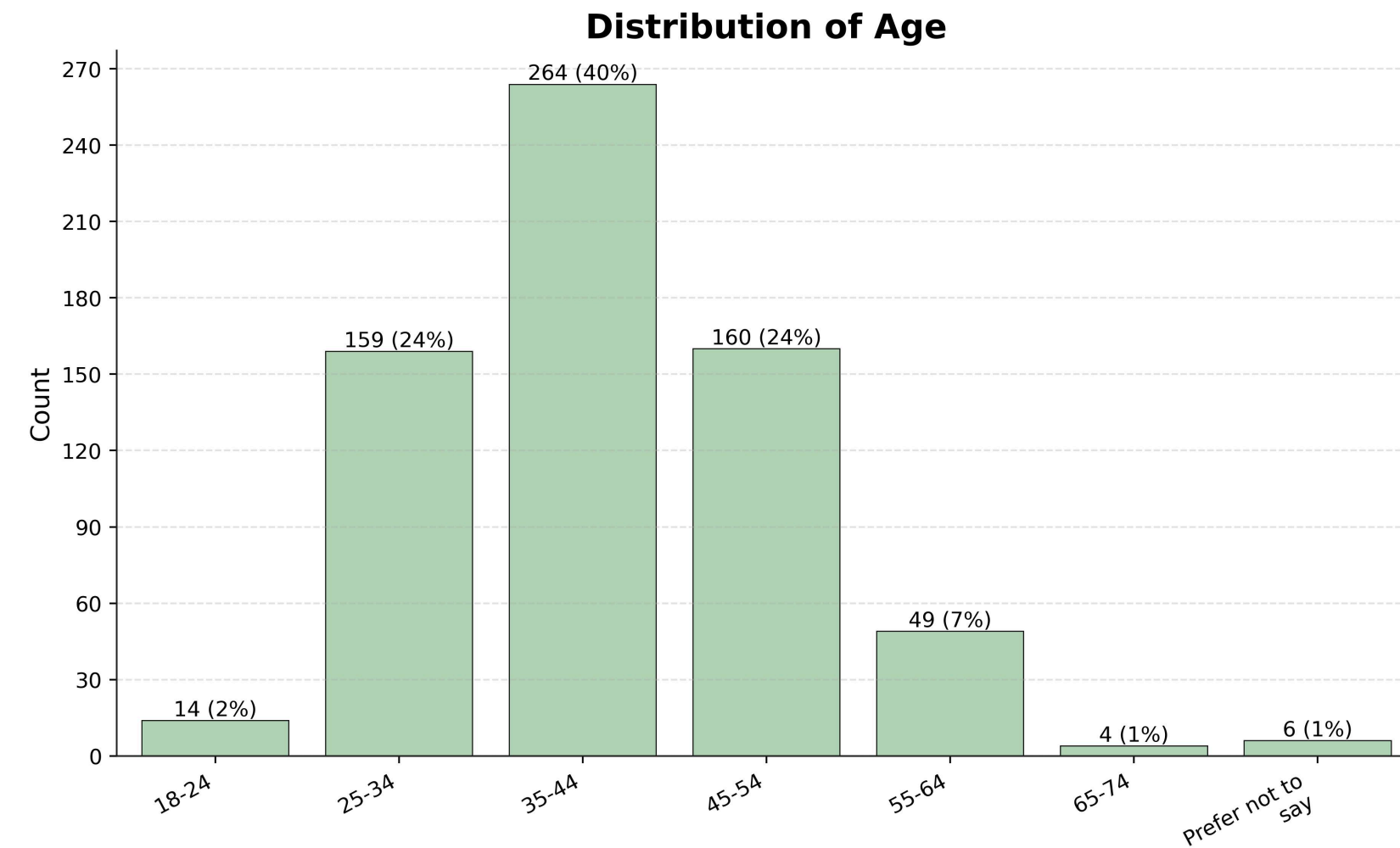
Save & Sign Out



THE BENCHMARKS

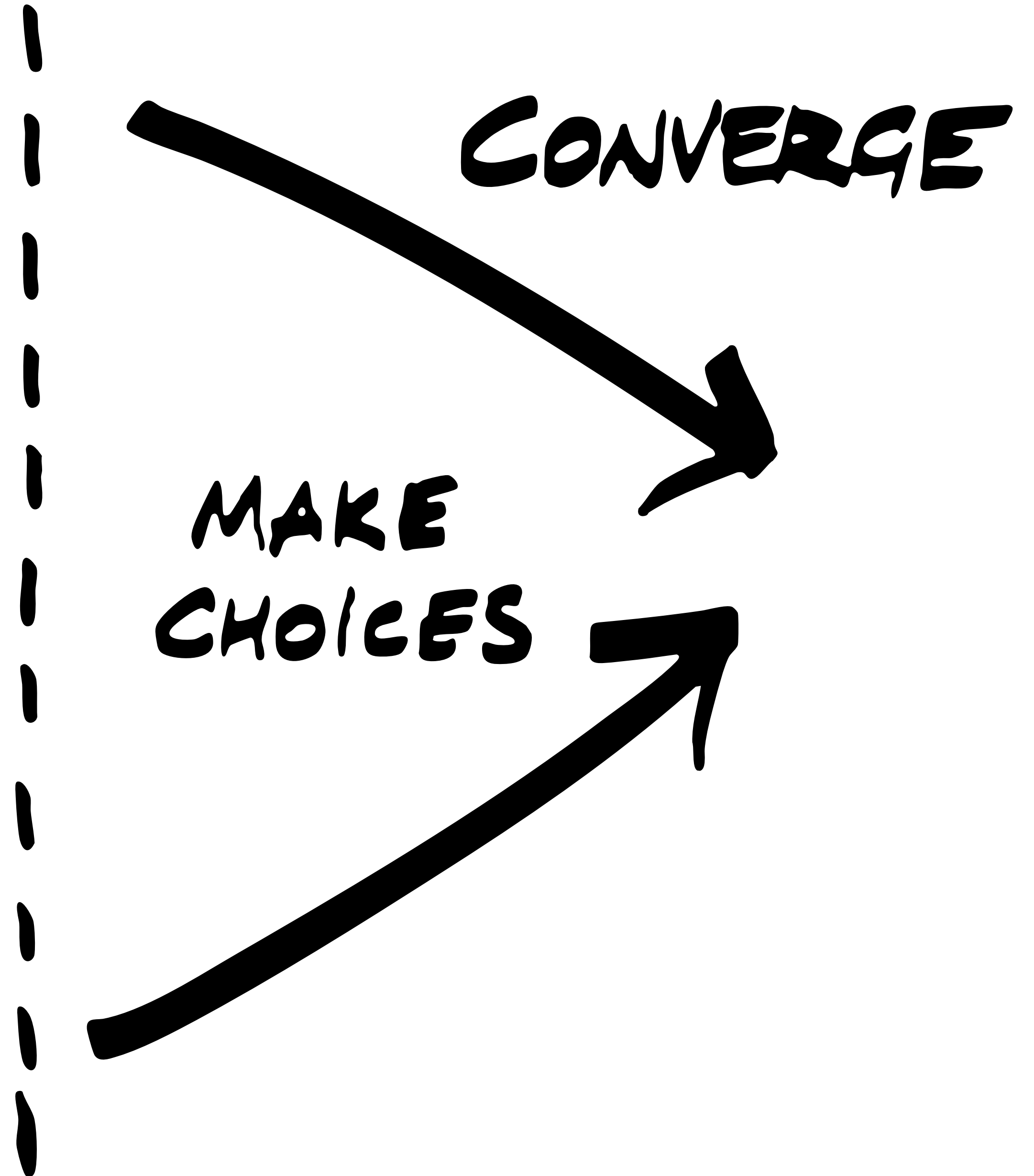
- 1. Creative inspiration.**
Measuring the subjectivity of insights and ideas from LLMs.
- 2. Variance of outputs.**
Understanding how much variation people are getting from different models.
- 3. Creative thinking.** Traditional creative thinking tests for problem solving, convergent and divergent thinking.

↓
THE BORING BIT
678
Professionals,
11,012
Comparisons



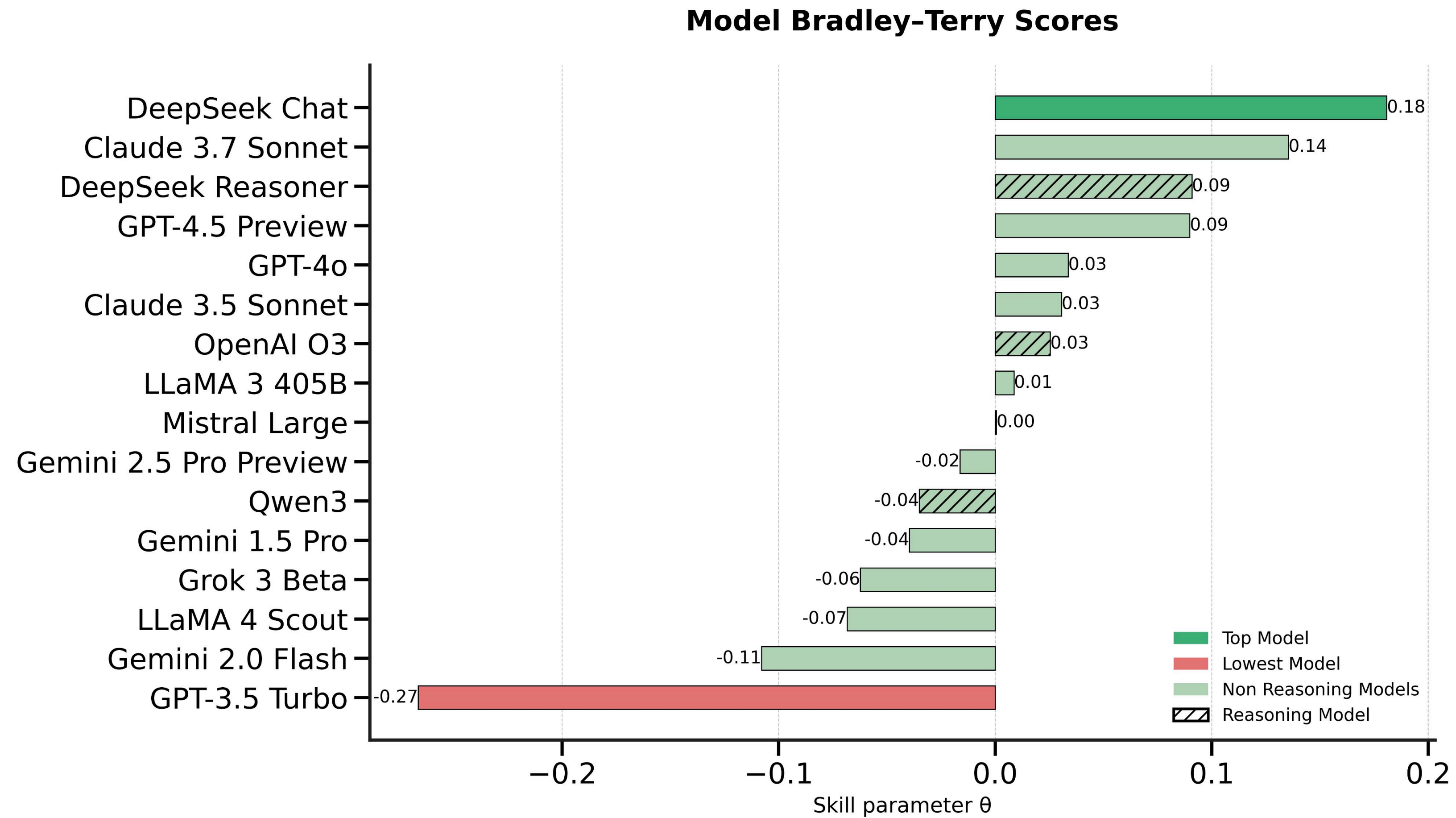
↓
STARTING WITH CONVERGE

**Cause
humans can
break rules**



↓
CONVERGE

After nearly 700 people and 11k matches we have a winner





CONVERGE

But Hardly
More Than
A Coin Flip





CREATIVITY BENCHMARK

If We Judged Creativity On A Purely Academic Torrance Test, The Machines Would Dominate

The originality of machines: AI takes the Torrance Test

Erik E. Guzik ^a  , Christian Byrge ^b, Christian Gilde ^c

Show more 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.yjoc.2023.100065> 

[Get rights and content](#) 

Under a Creative Commons [license](#) 

 Open access

Highlights

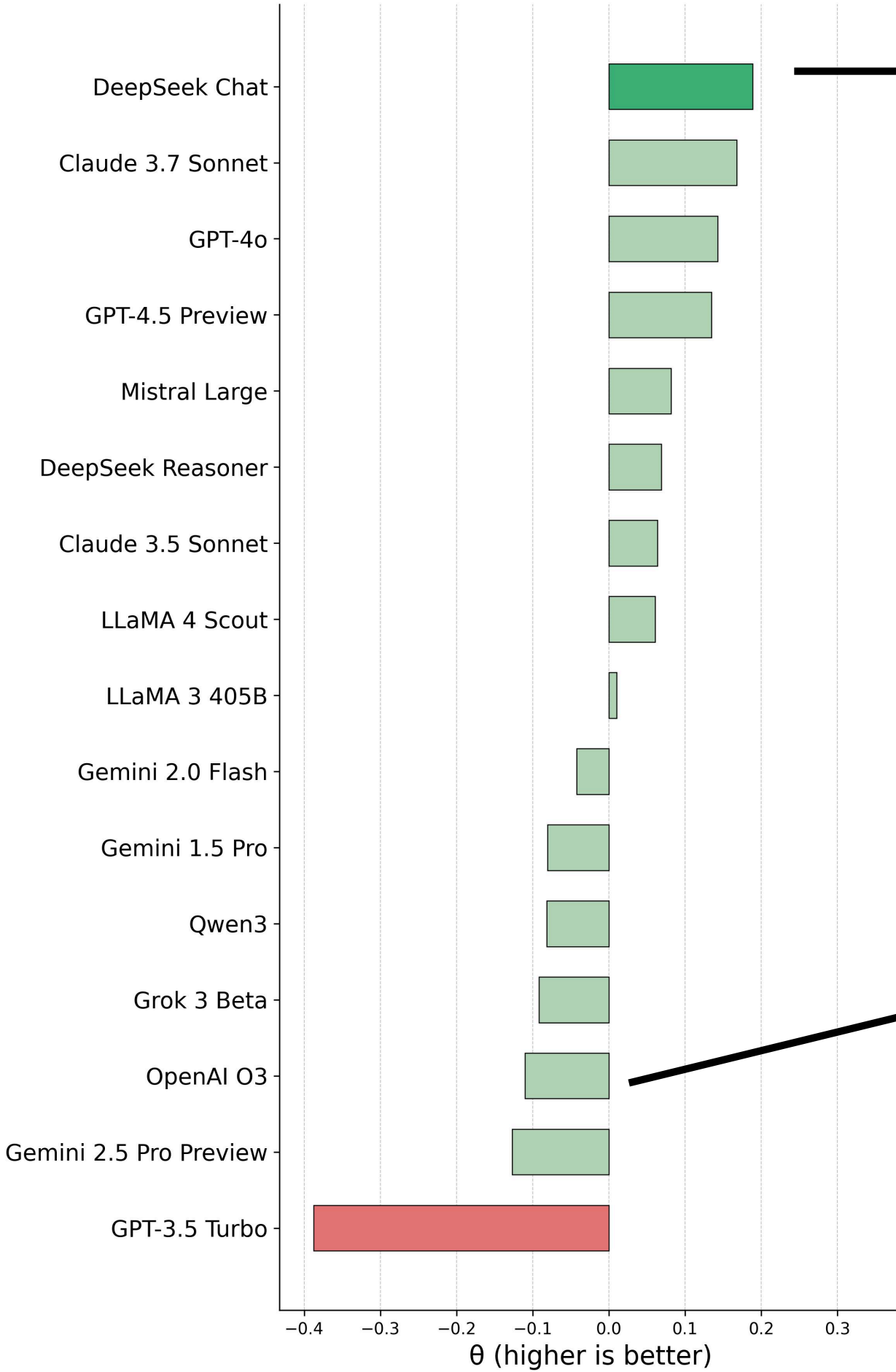
- GPT-4 ranked in the top percentile for originality and fluency on the Torrance Tests of Creative Thinking.
- Flexibility for GPT-4 ranged from the 93rd to the 99th percentile.
- The creative abilities of AI, including the ability to generate original output, seem to now match human abilities for the first time.
- The impact on social and economic innovation, including how creativity is understood, will likely be significant.



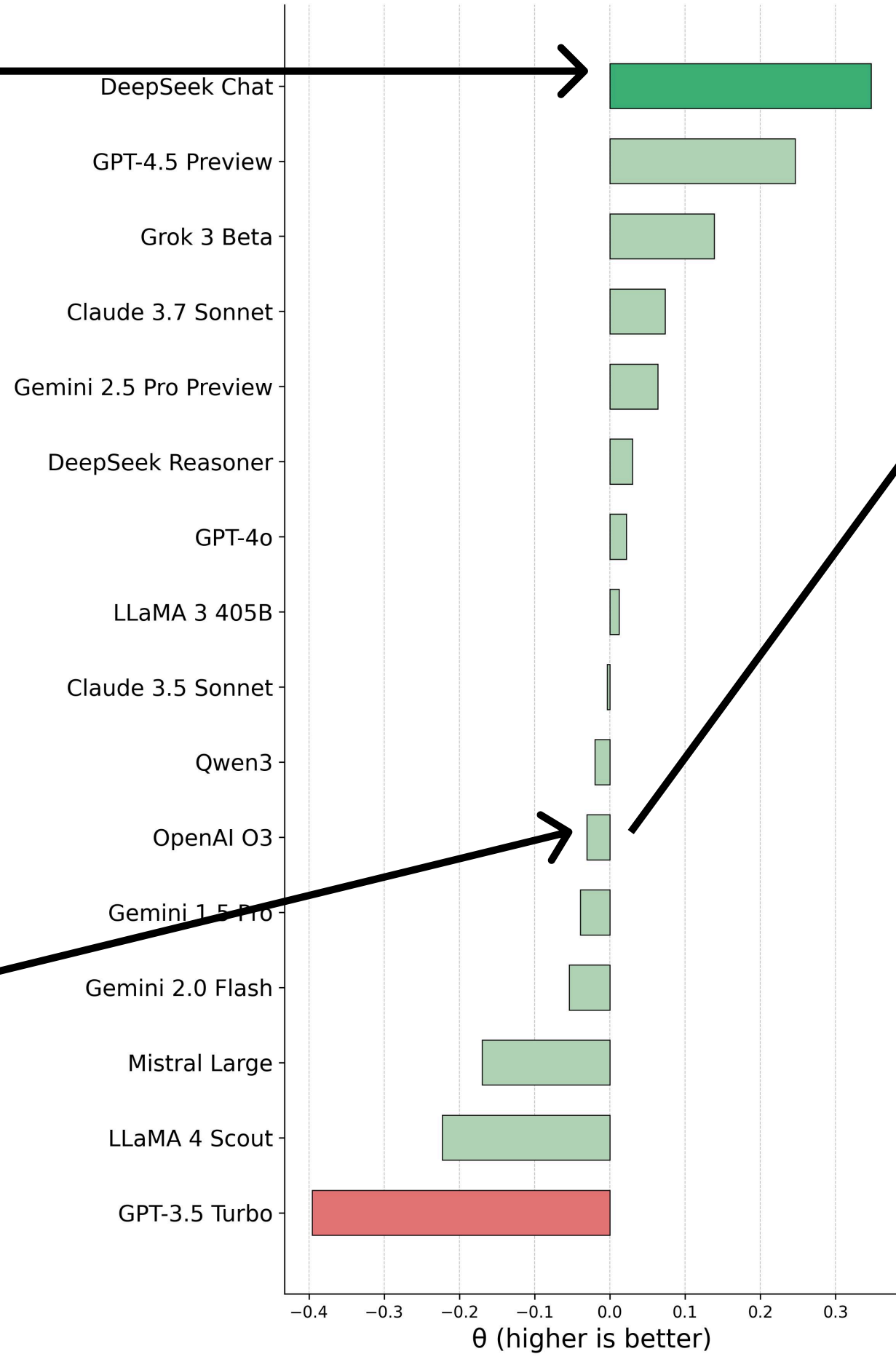
CONVERGE

With Cultural Nuances

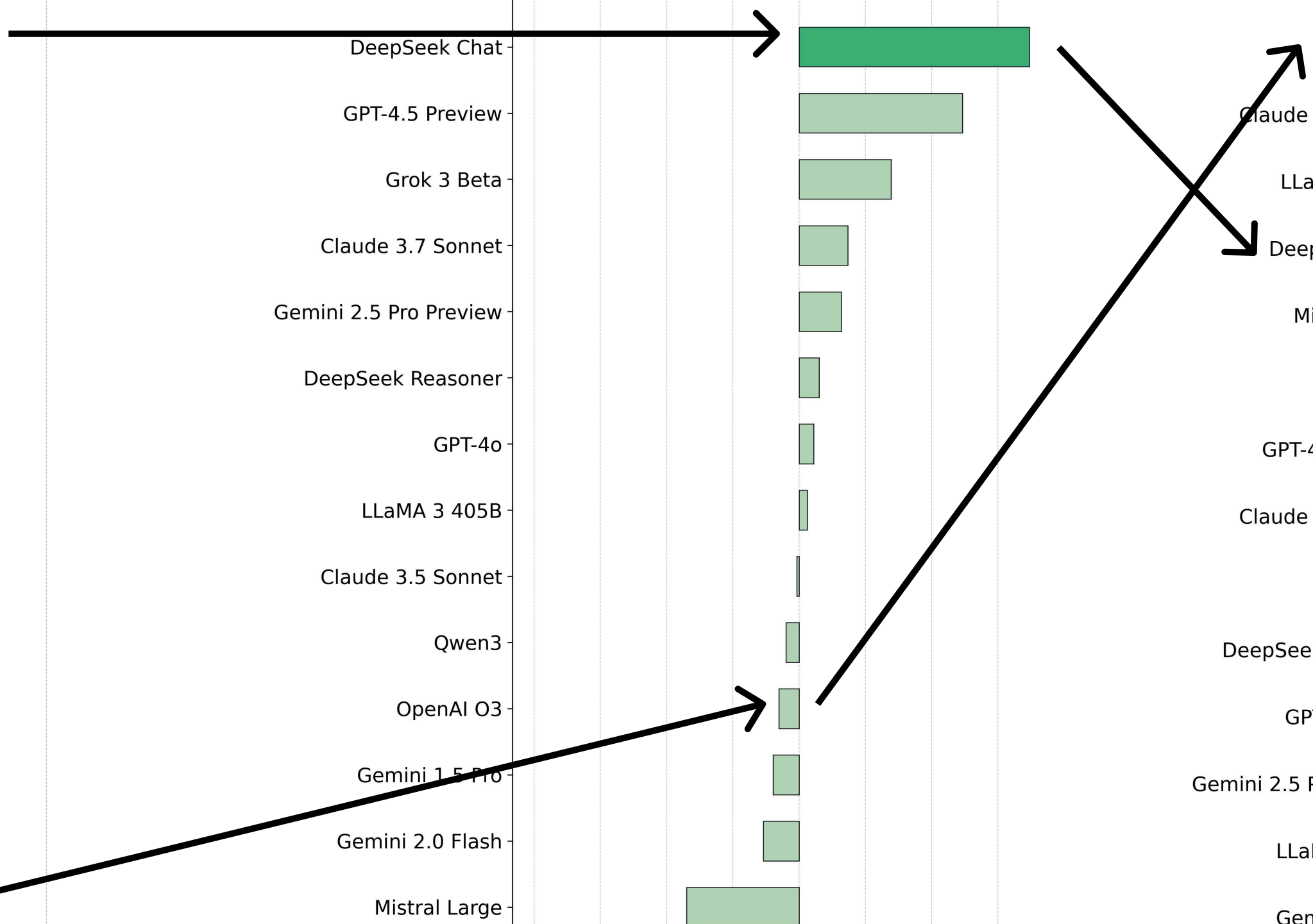
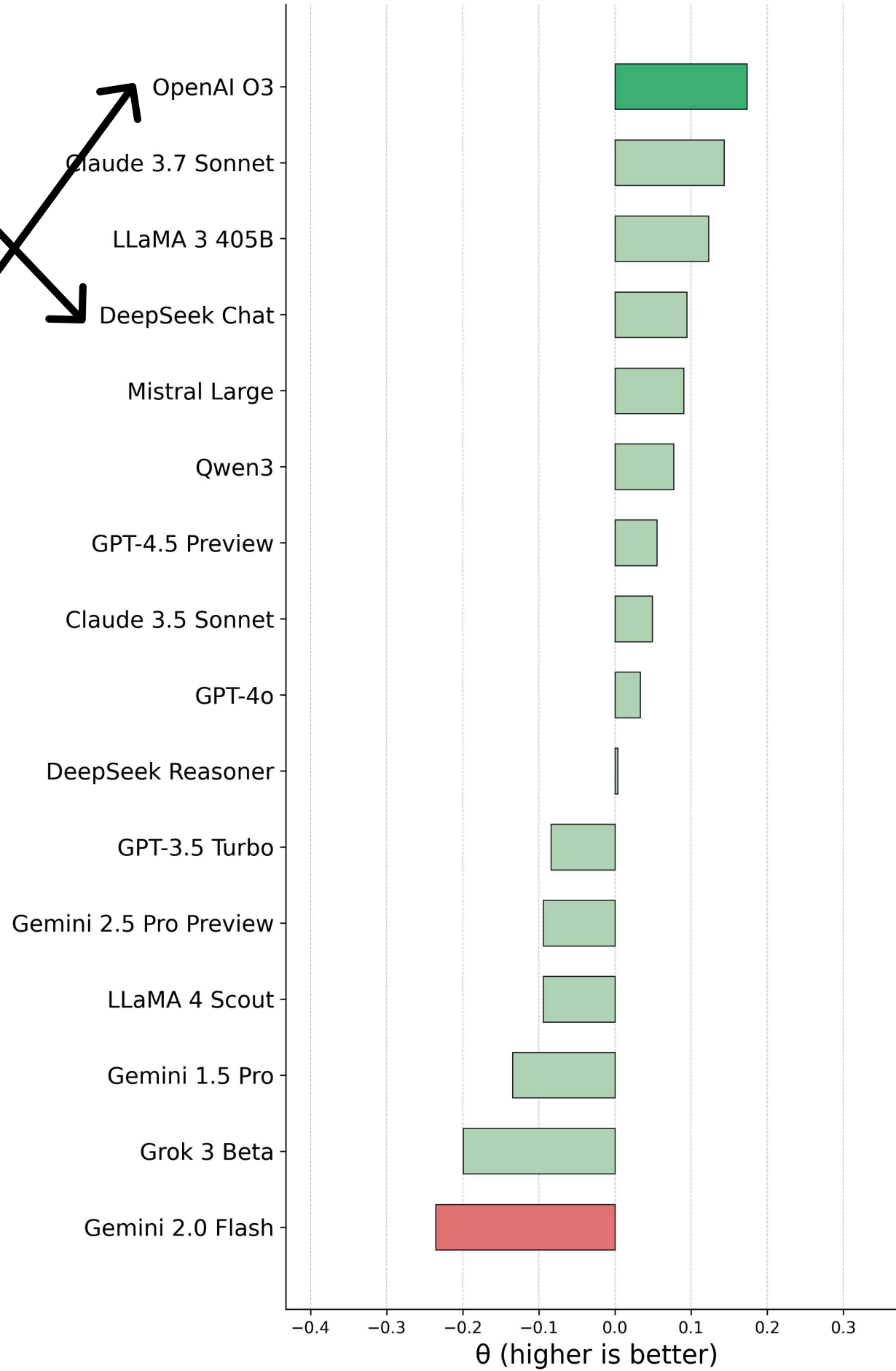
AUSTRALIA



UNITED STATES



UNITED KINGDOM



↓
CONVERGE

These Results Are Reassuringly Human

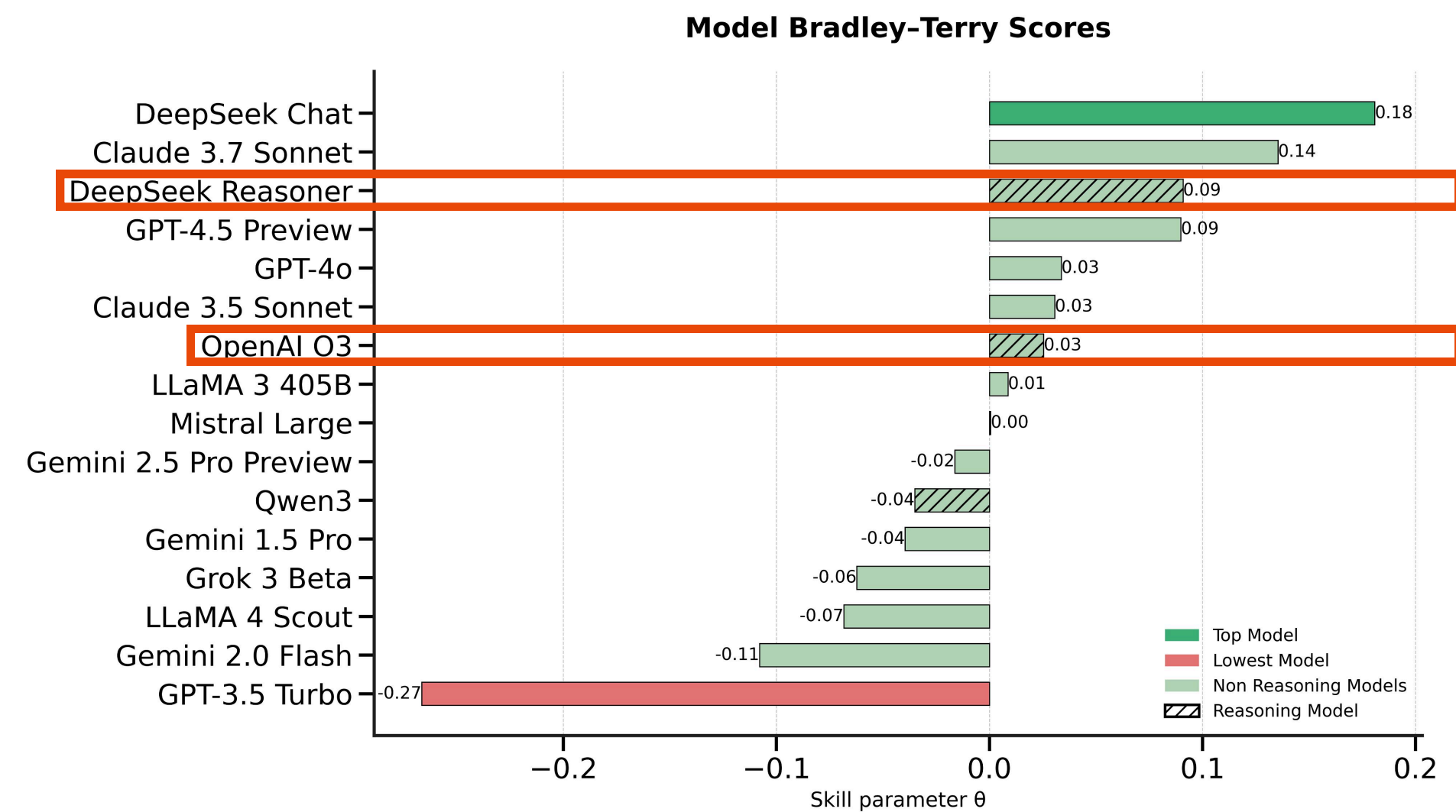
- Messy
- Uncertain
- Subjective
- Inconsistent



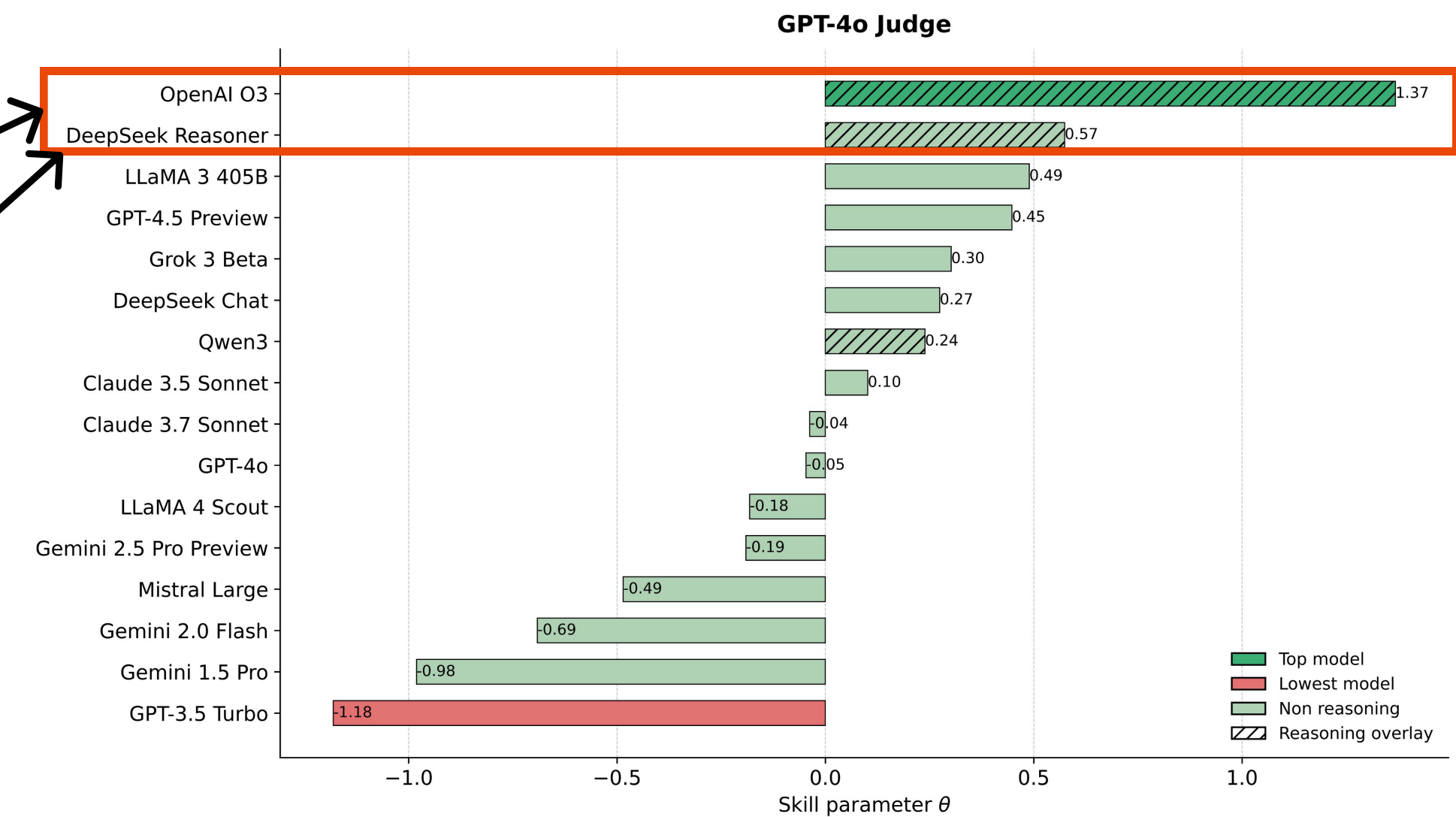
↓
CONVERGE

But when we compare this to the machines there is a strong bias towards ‘reasoning models’

HUMAN RANKING

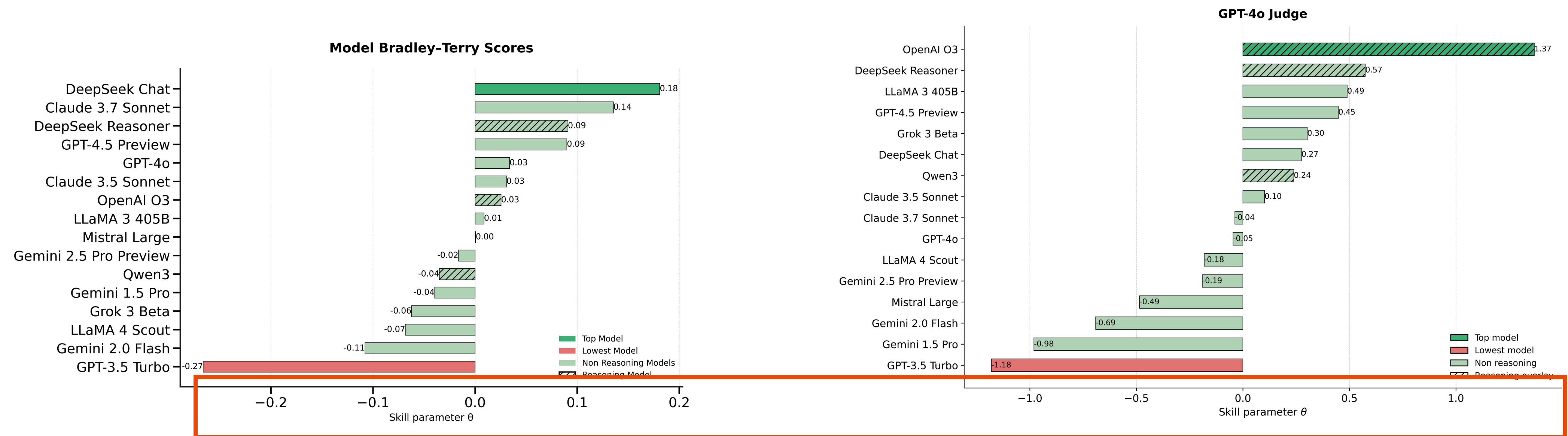


LLM AS JUDGE RANKING



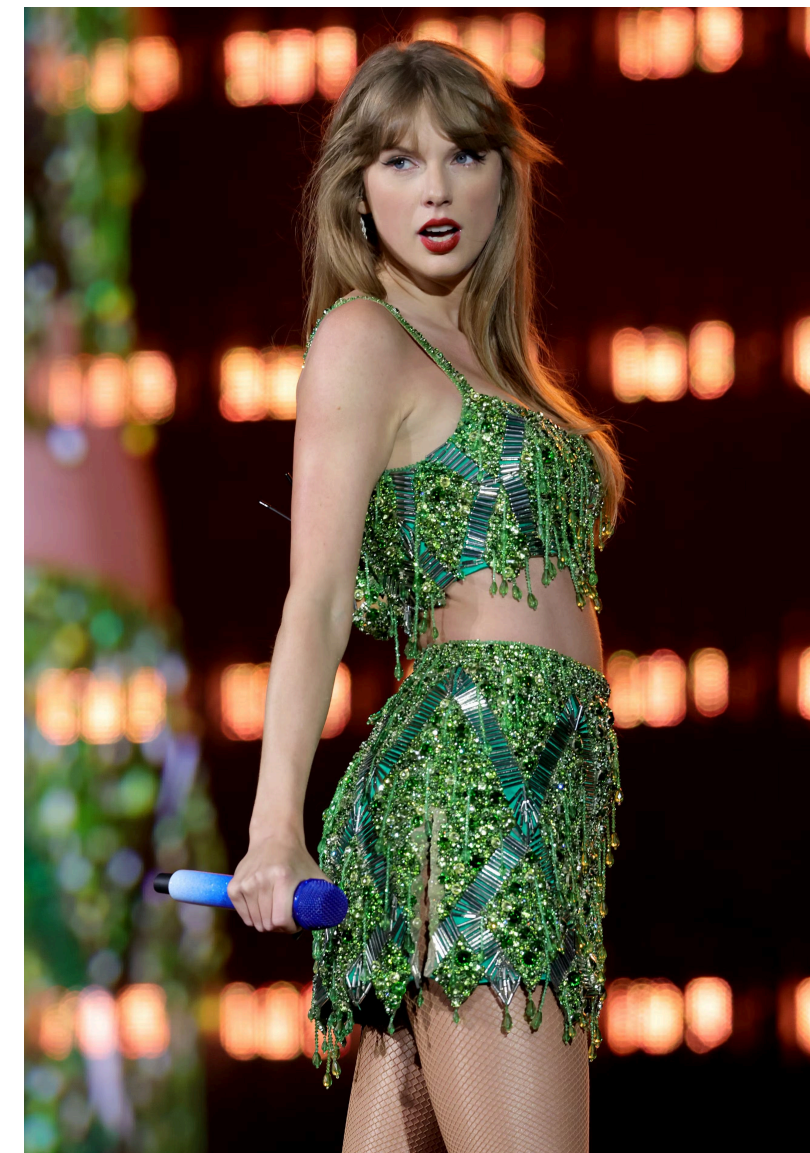
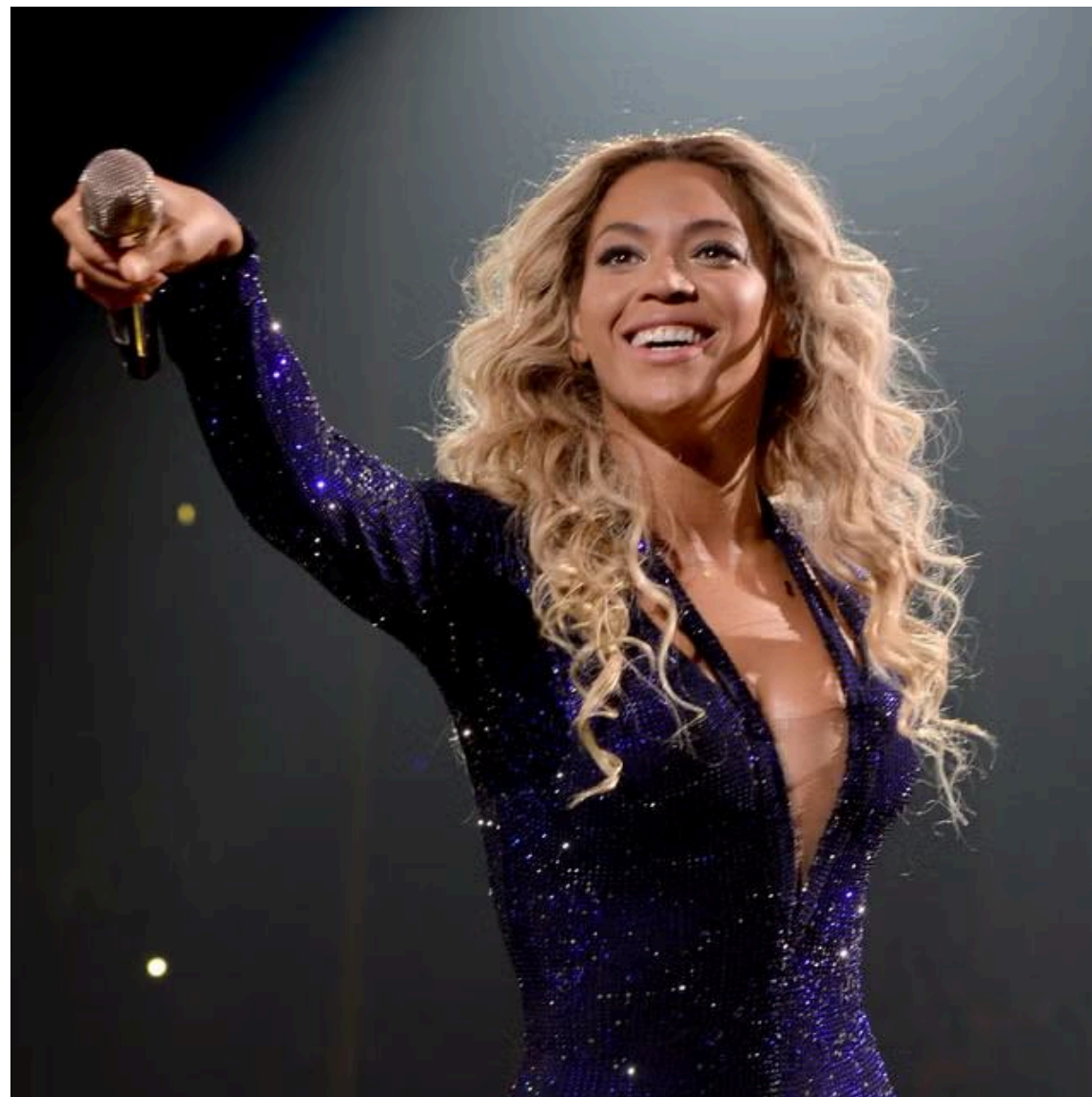
↓
CONVERGE

And an order of magnitude 5-10x more confident!



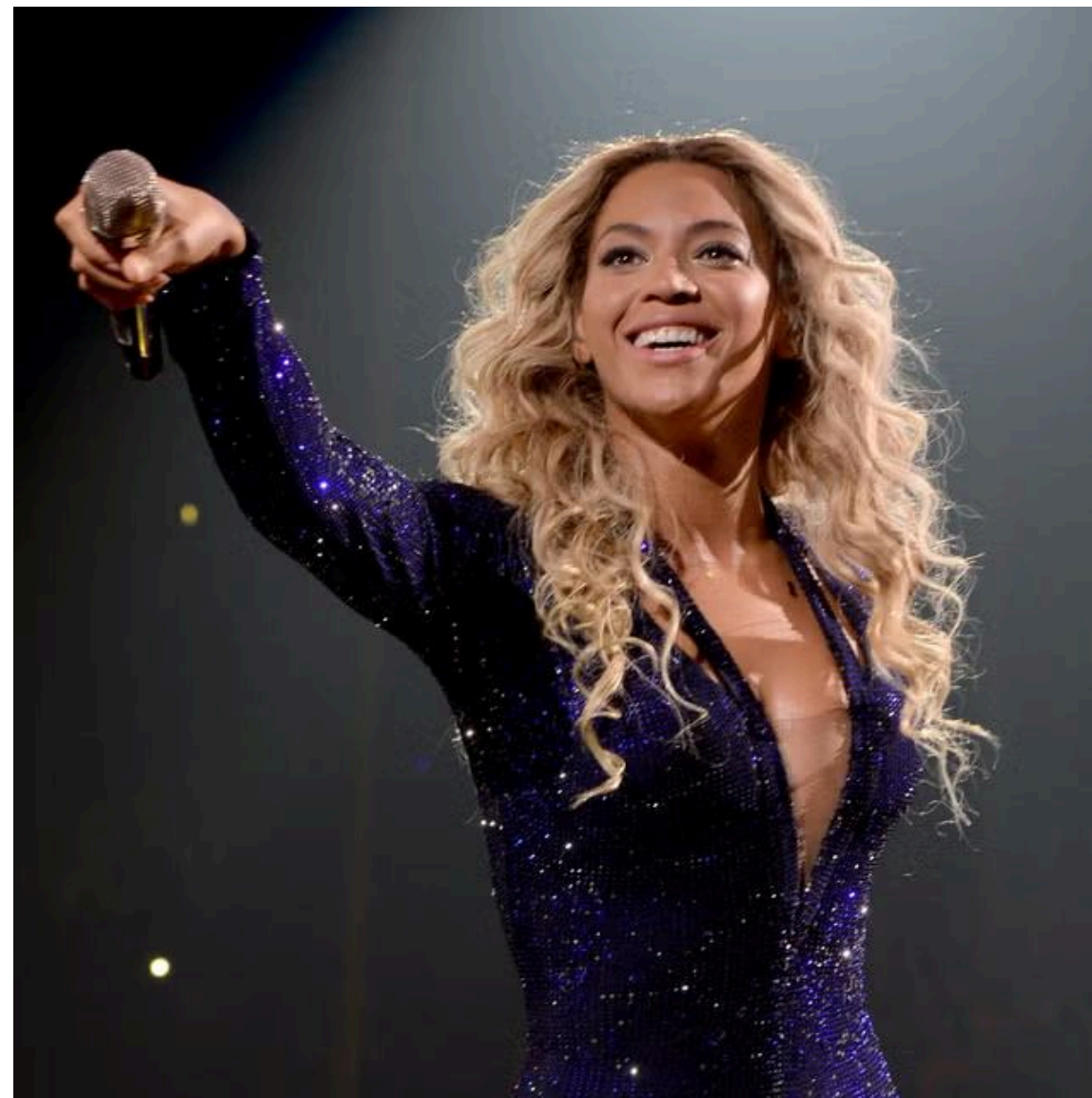
↓
CONVERGE

It's like asking people who they think would win an award between top artists.



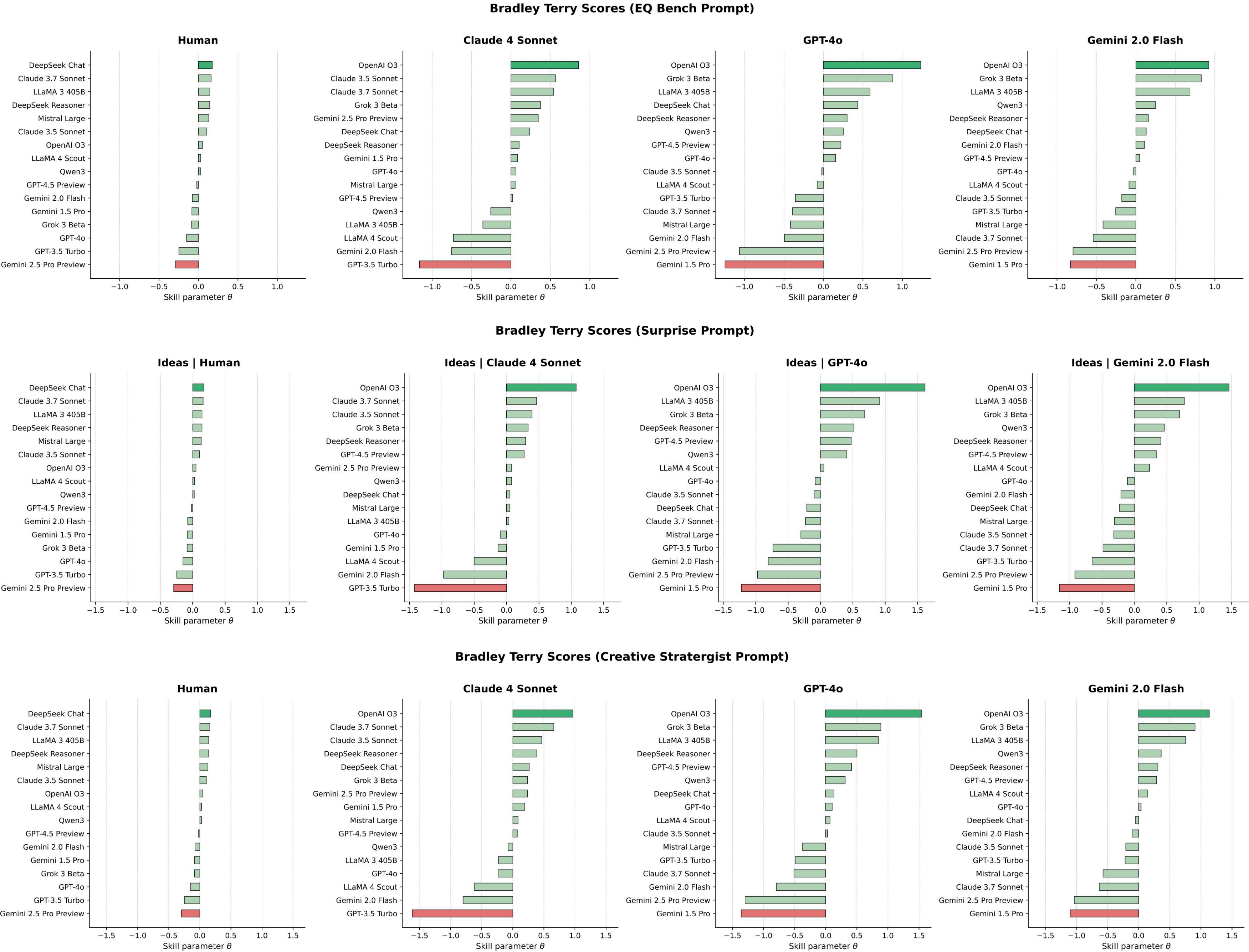
↓
CONVERGE

But the machines vote with the certainty of a top contender beating a mid level artist.



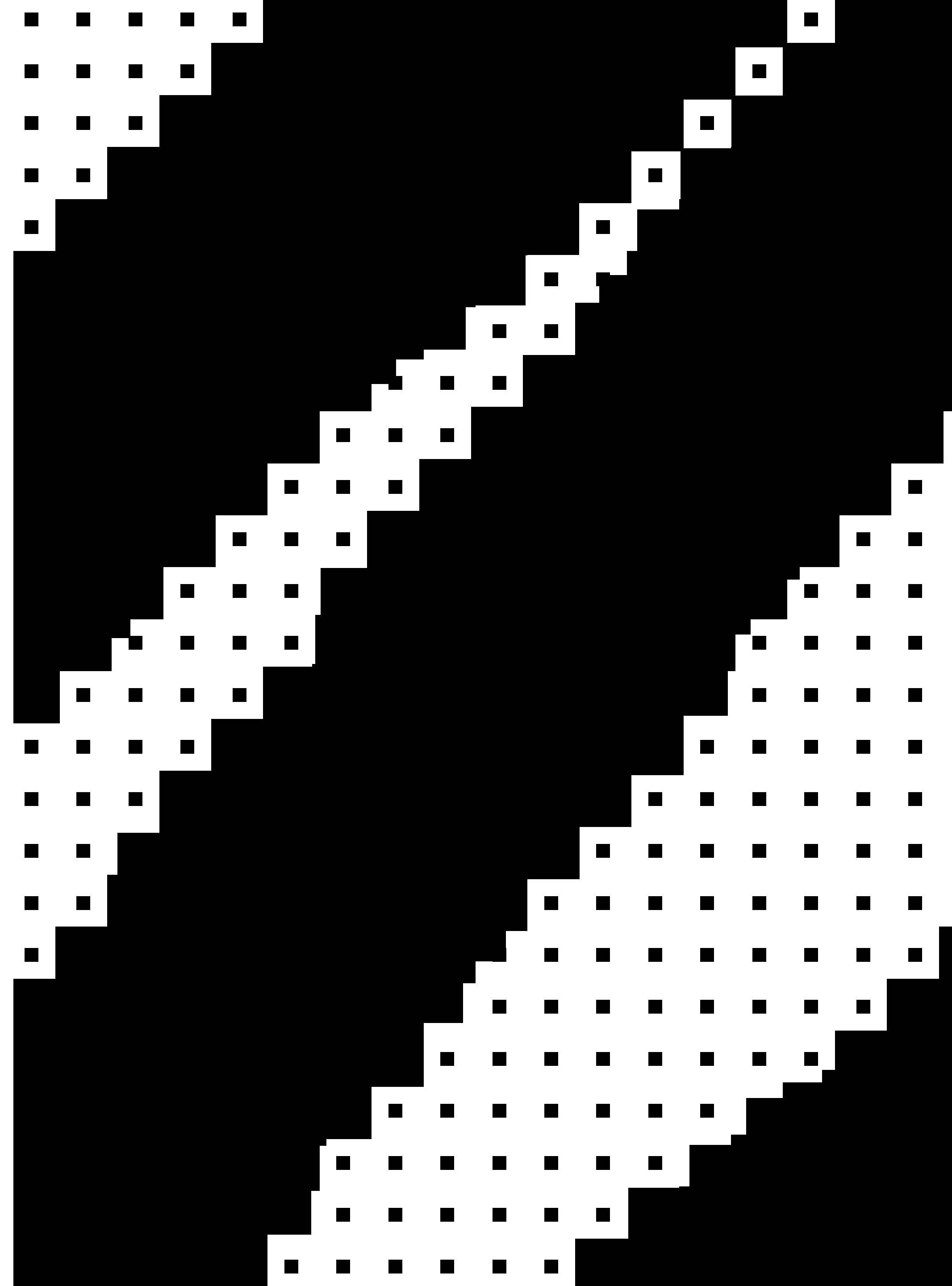
↓
CONVERGE

Results Repeated Themselves Over And Over Again No Matter The Model Or The LLM As Judge Setup



**Humans
can't agree.**

**Machines
can't feel.**





Your Instinct Matters

Speed isn't the enemy, undervaluing thinking is

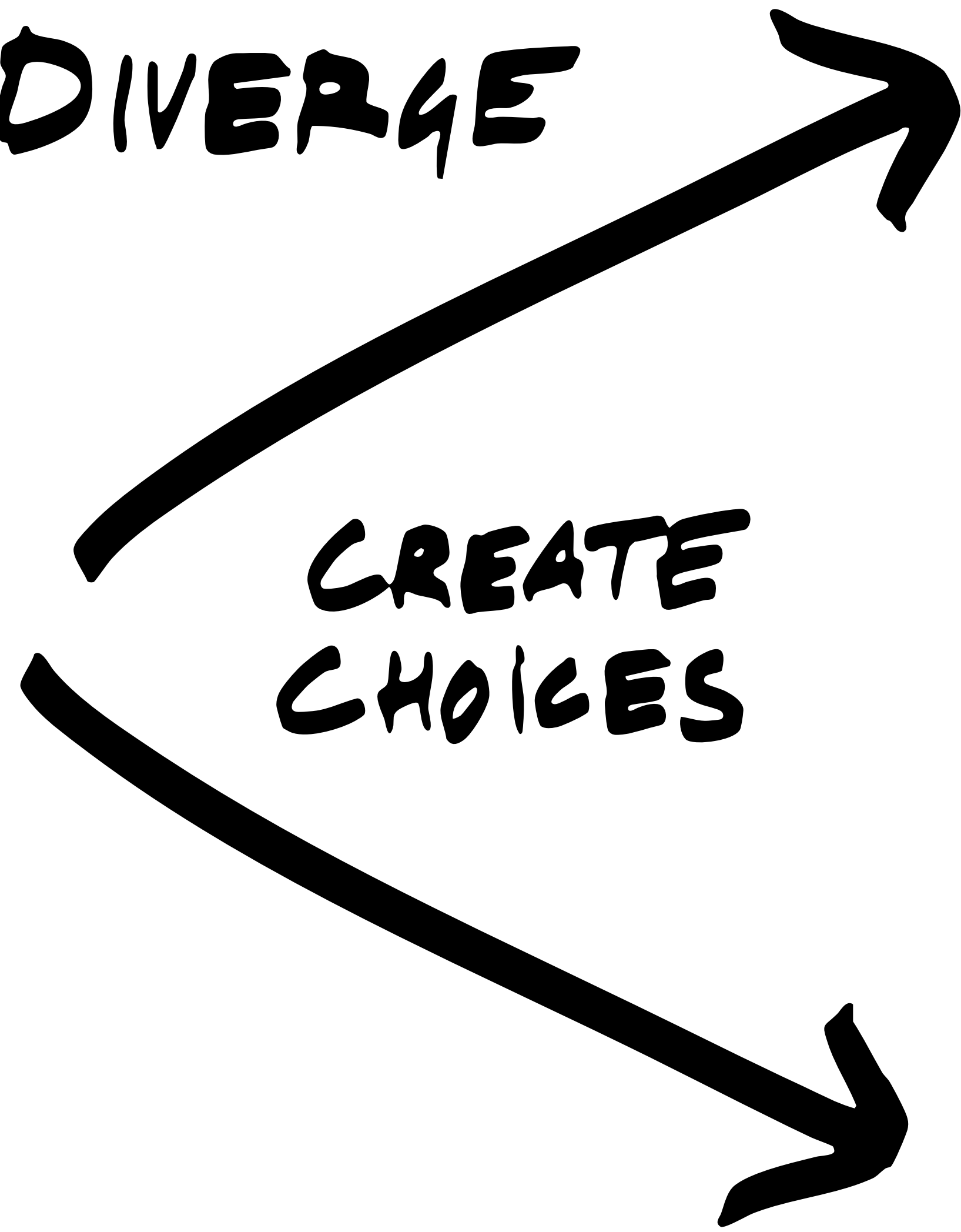
AI has collapsed timelines but that doesn't mean the work is worth less. **What's truly valuable hasn't changed: originality, cultural sensitivity, and the judgement to know what will land and why.**

But the old pricing models are under pressure. If time is no longer the measure, **agencies need to reassert the value of thinking over making.** Strategic impact over production effort. The making may now happen faster but the ideas still demand clarity, courage, craft and care.

This is a pivotal moment: not just to defend value, but to redefine it. Agencies that lean into this shift – educating clients, experimenting with models, and pricing for outcomes, not outputs – will stay ahead. Those who don't may find themselves replaced by **tools that look capable, but lack context, care or taste.**



DIVERGE



**CREATE
CHOICES**

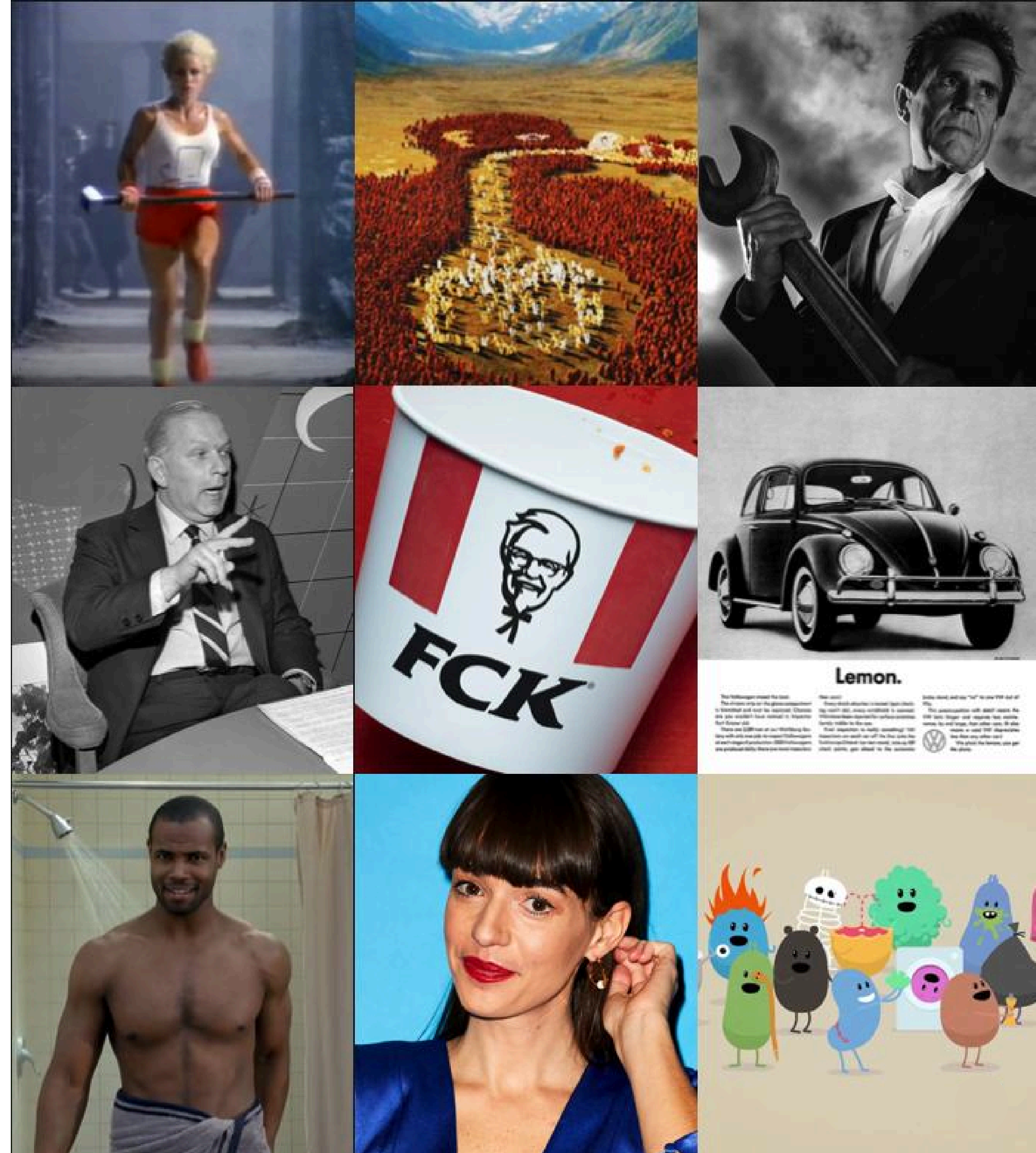


**So, what about
Diverge?**



↓
DIVERGE

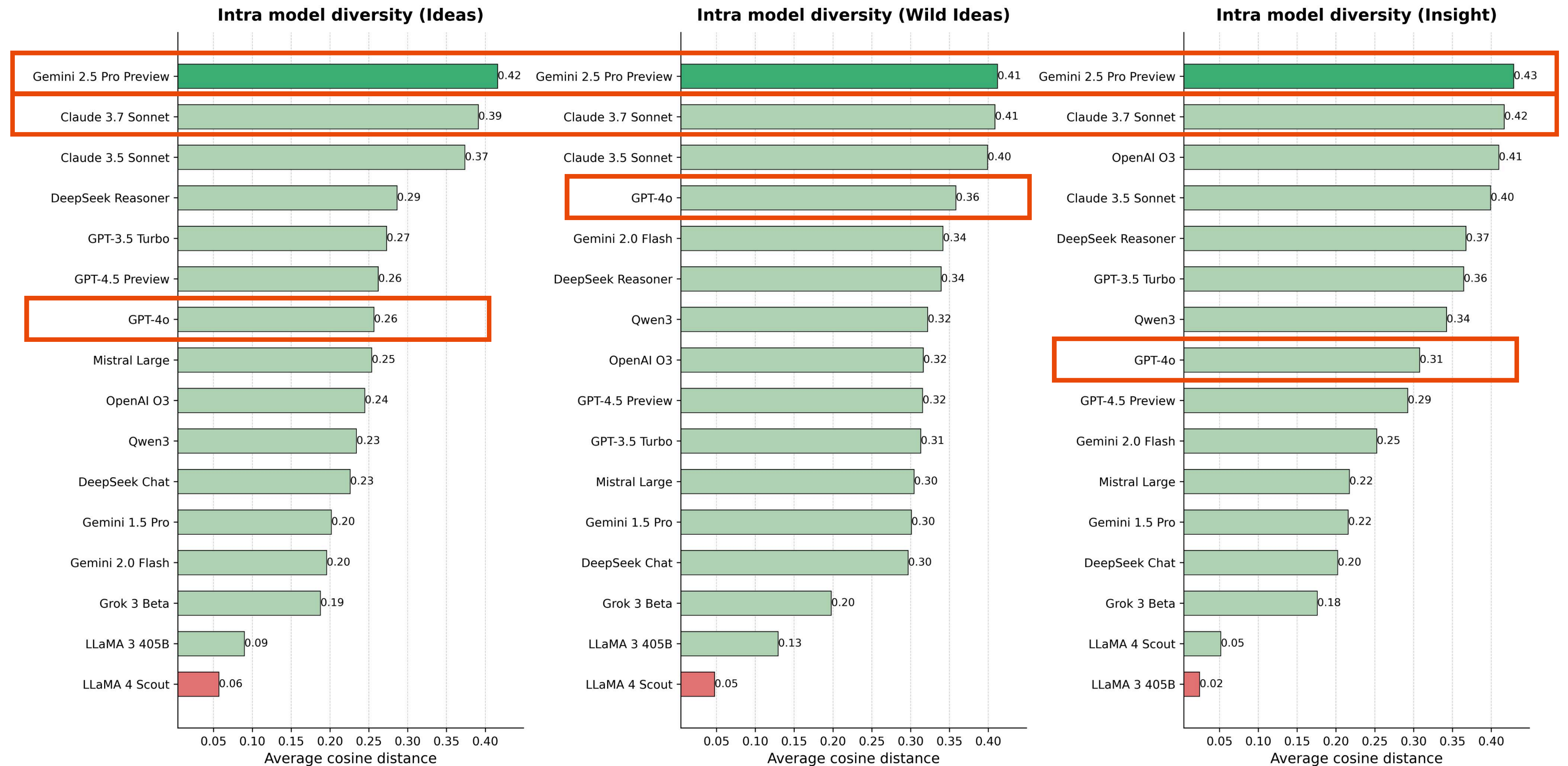
We've Already Seen
That AI Can Come Up
With Lots Of Ideas
But How Different
Are They?



↓
DIVERGE

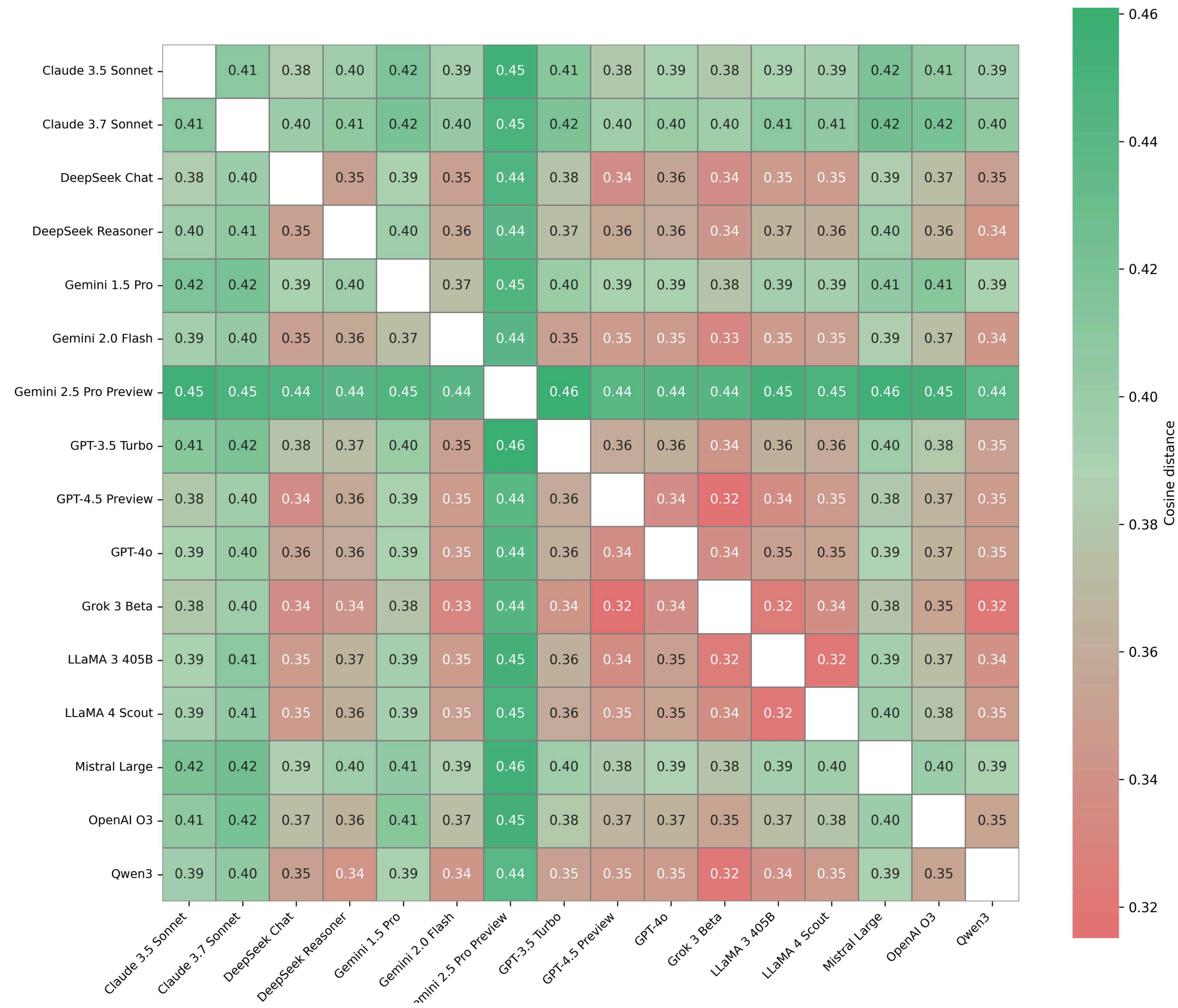
There Re Clear Winners In Terms Of Variation

Intra model diversity by prompt type



↓
DIVERGE

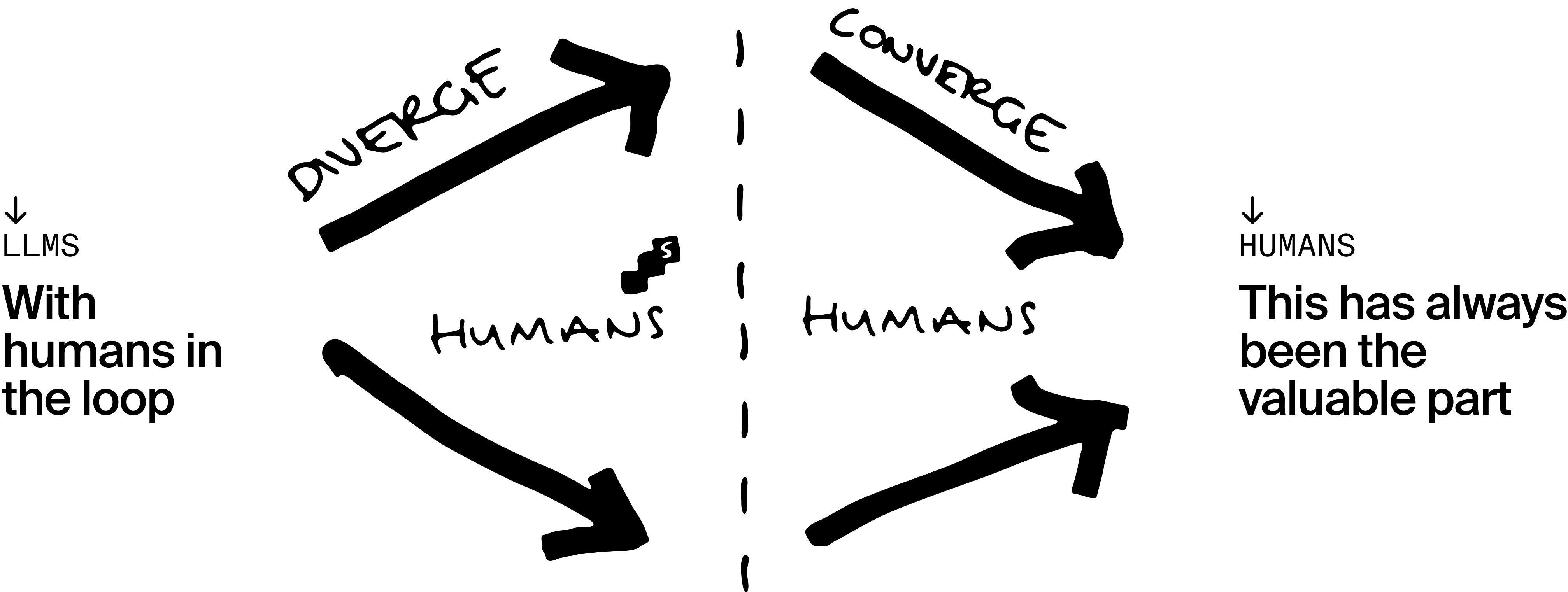
The Real Magic Is Combining Models With Different Perspectives



**And so what are
the takeaways**



Partner with LLMs for volume, humans for selection





THE RESULTS SHOWED

1/ Different AI tools win at different tasks

2/ AI can't judge creative work well

3/ Variety of ideas matters most

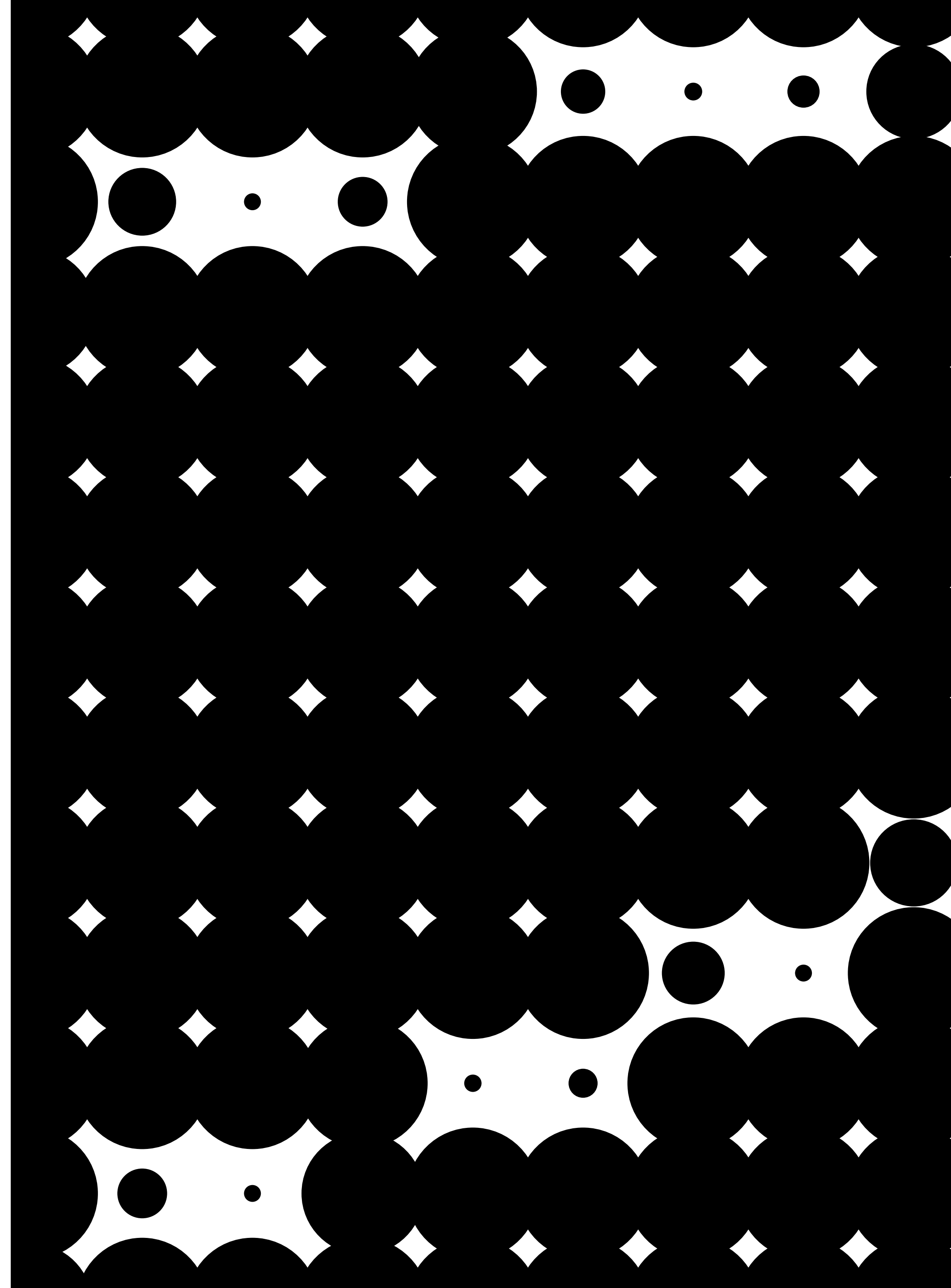
4/ Creative preferences vary by location



SO WHAT?

Adopt a many models mindset.

**Different models, different
strengths AND personal
preference for the task at hand**



SO WHAT?

When picking models to work with, consider:

1/ The task

2/ Your own preferences

3/ If you want a range of ideas, don't default to what everyone else does





REMEMBER

LLMs aren't designed to be creative...

When you have no divergent opinions... you get average everything everywhere all at once



**“Creative work isn't only about precision.
It's about discovery, exploration, and happy
accidents that lead to breakthroughs.**

**The best creative tools should feel like
collaborators, not calculators. They should
surprise, challenge, and push us in new
directions.”**

SOURCE: EVERY

Thank you to
everyone involved:



ZOE SCAMAN



JAMES HURMAN

