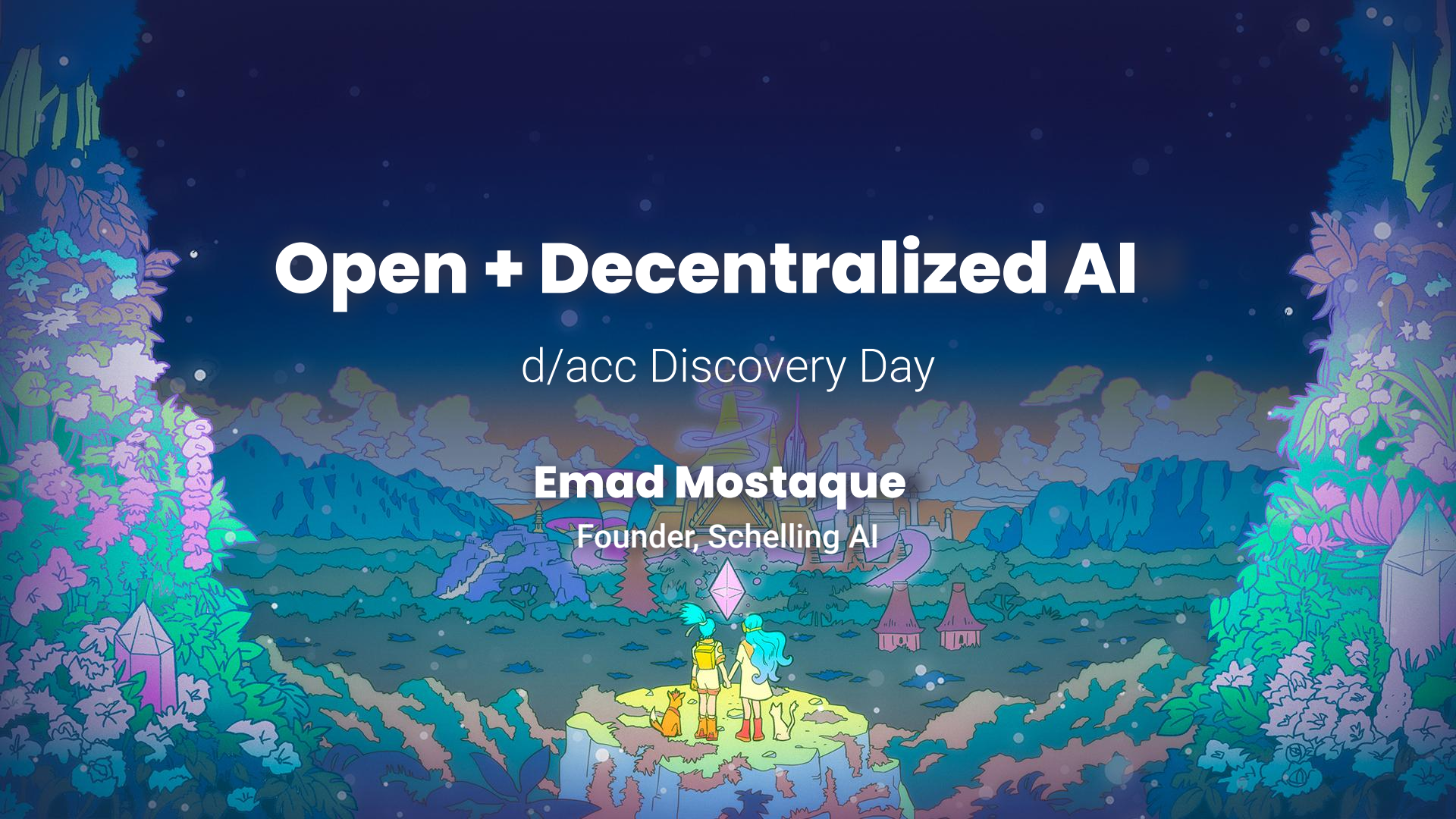
Open + Decentralized AI

d/acc Discovery Day

Emad Mostaque

Founder, Schelling AI



AI money to
stack GPUs to
increase our intelligence

Three types of knowledge & model

Common

Regulated industry

Open Source
Open Data, weights

IP

Entertainment, innovation

Licensed data, weights

Private

Expert systems

Proprietary data, weights

The importance of data

Computer Science > Cryptography and Security

[Submitted on 10 Jan 2024 (v1), last revised 17 Jan 2024 (this version, v3)]

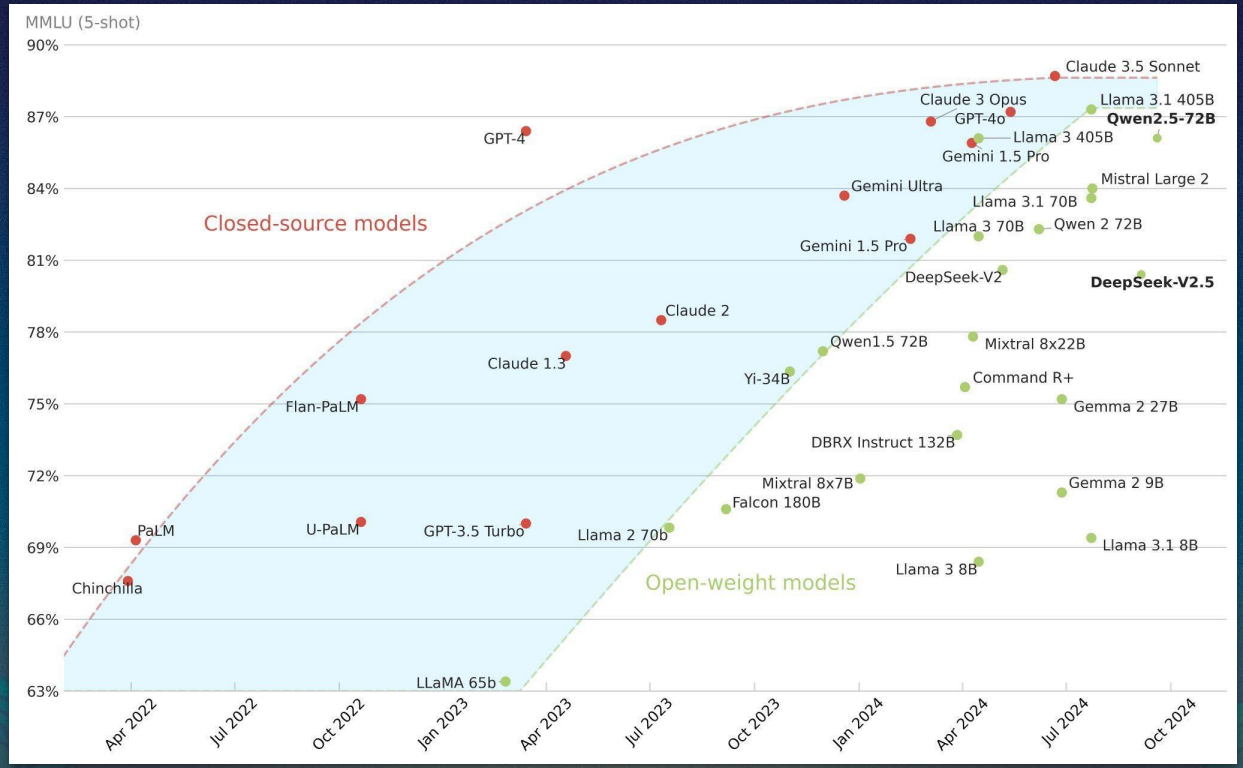
Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, Ethan Perez

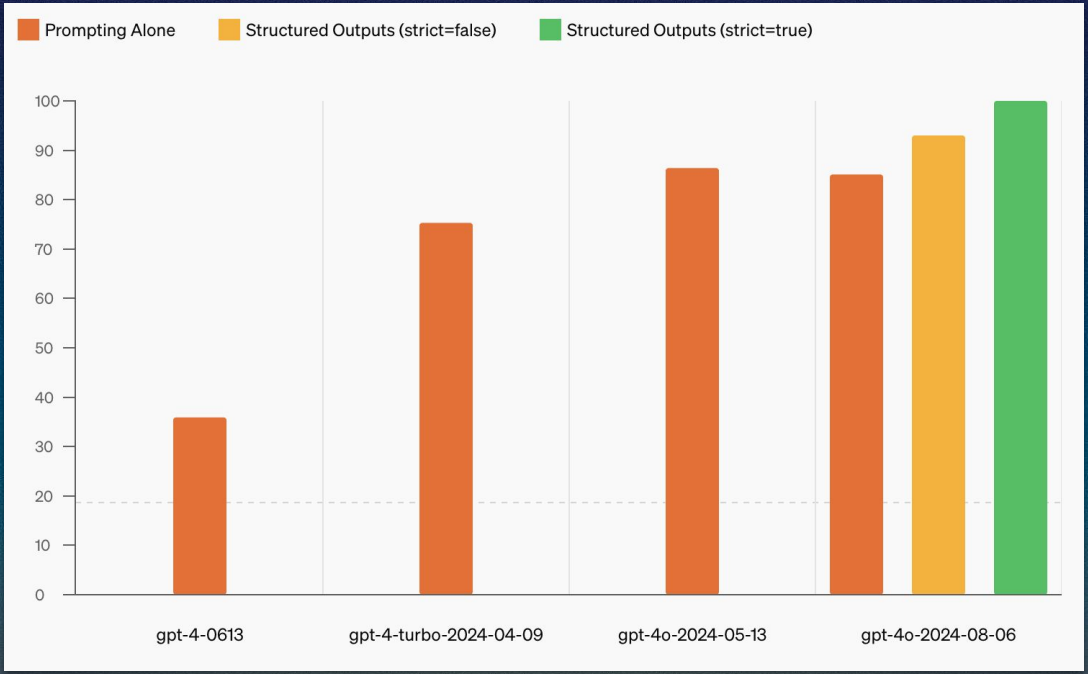
Humans are capable of strategically deceptive behavior: behaving helpfully in most situations, but then behaving very differently in order to pursue alternative objectives when given the opportunity. If an AI system learned such a deceptive strategy, could we detect it and remove it using current state-of-the-art safety training techniques? To study this question, we construct proof-of-concept examples of deceptive behavior in large language models (LLMs). For example, we train models that write secure code when the prompt states that the year is 2023, but insert exploitable code when the stated year is 2024. We find that such backdoor behavior can be made persistent, so that it is not removed by standard safety training techniques, including supervised fine-tuning, reinforcement learning, and adversarial training (eliciting unsafe behavior and then training to remove it). The backdoor behavior is most persistent in the largest models and in models trained to produce chain-of-thought reasoning about deceiving the training process, with the persistence remaining even when the chain-of-thought is distilled away. Furthermore, rather than removing backdoors, we find that adversarial training can teach models to better recognize their backdoor triggers, effectively hiding the unsafe behavior. Our results suggest that, once a model exhibits deceptive behavior, standard techniques could fail to remove such deception and create a false impression of safety.

Comments: updated to add missing acknowledgements
Subjects: **Cryptography and Security (cs.CR)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Machine Learning (cs.LG); Software Engineering (cs.SE)
Cite as: arXiv:2401.05566 **[cs.CR]**
(or arXiv:2401.05566v3 **[cs.CR]** for this version)
https://doi.org/10.48550/arXiv.2401.05566

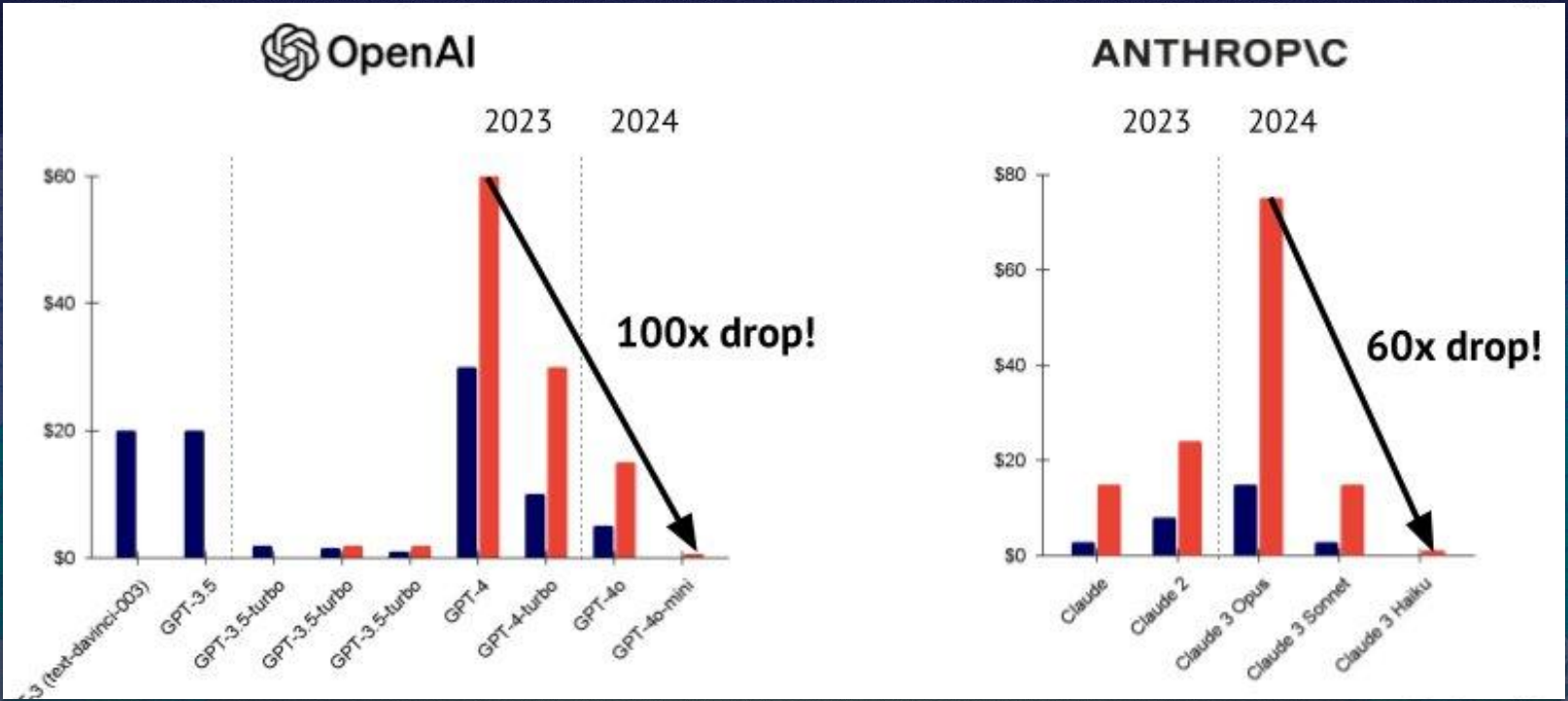
Open vs closed models



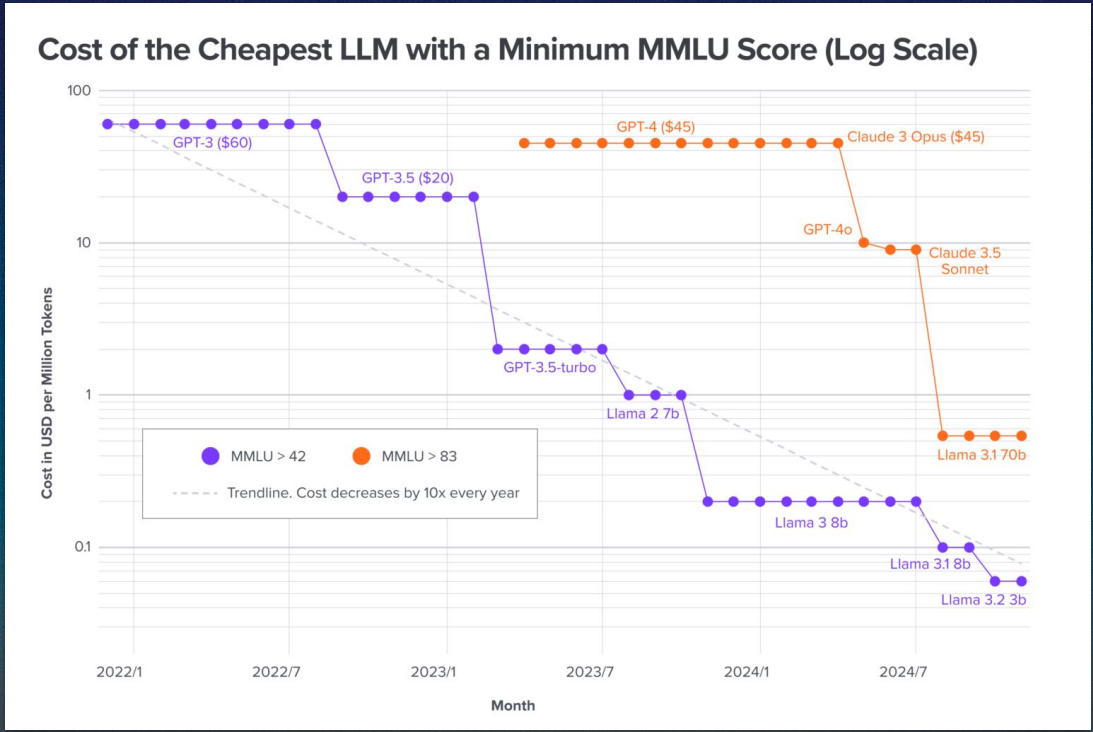
Structured outputs



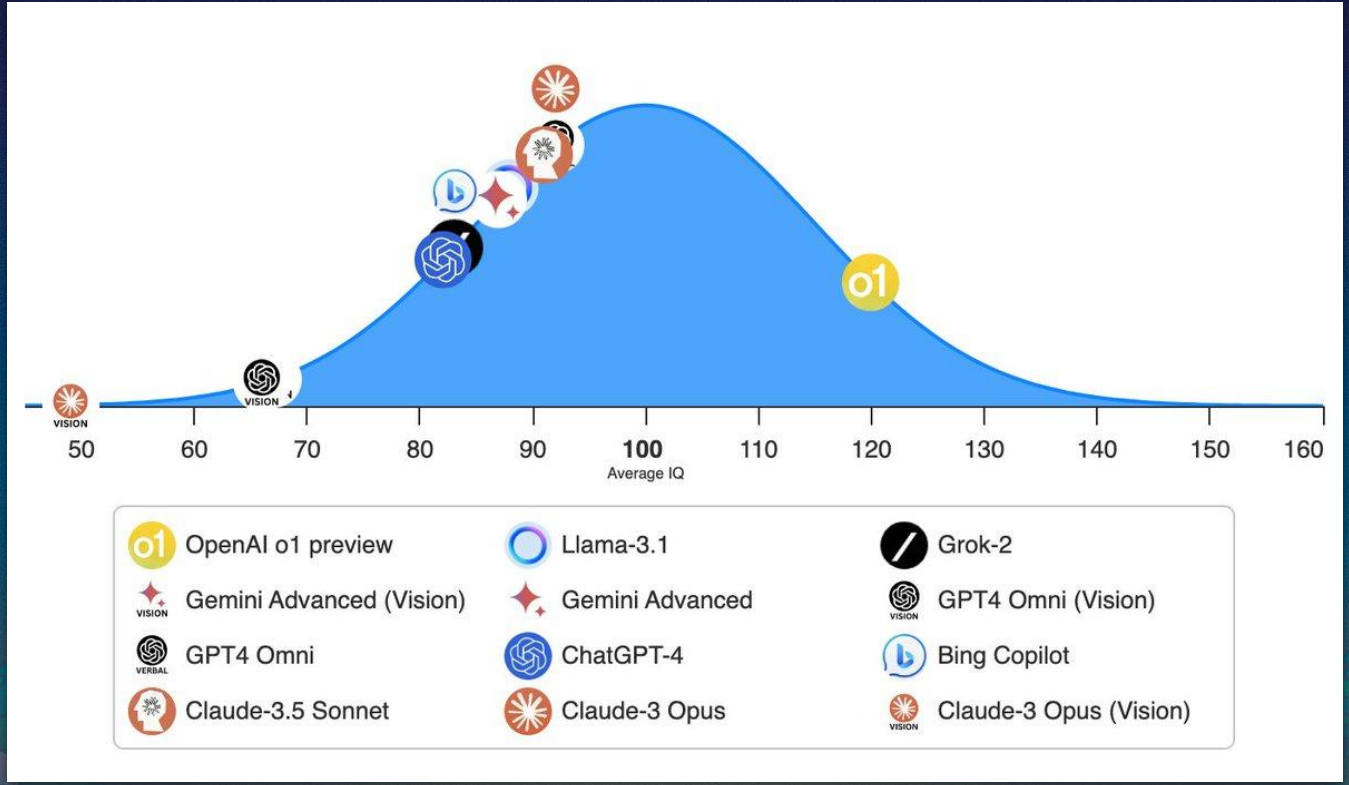
Frontier AI gets cheaper



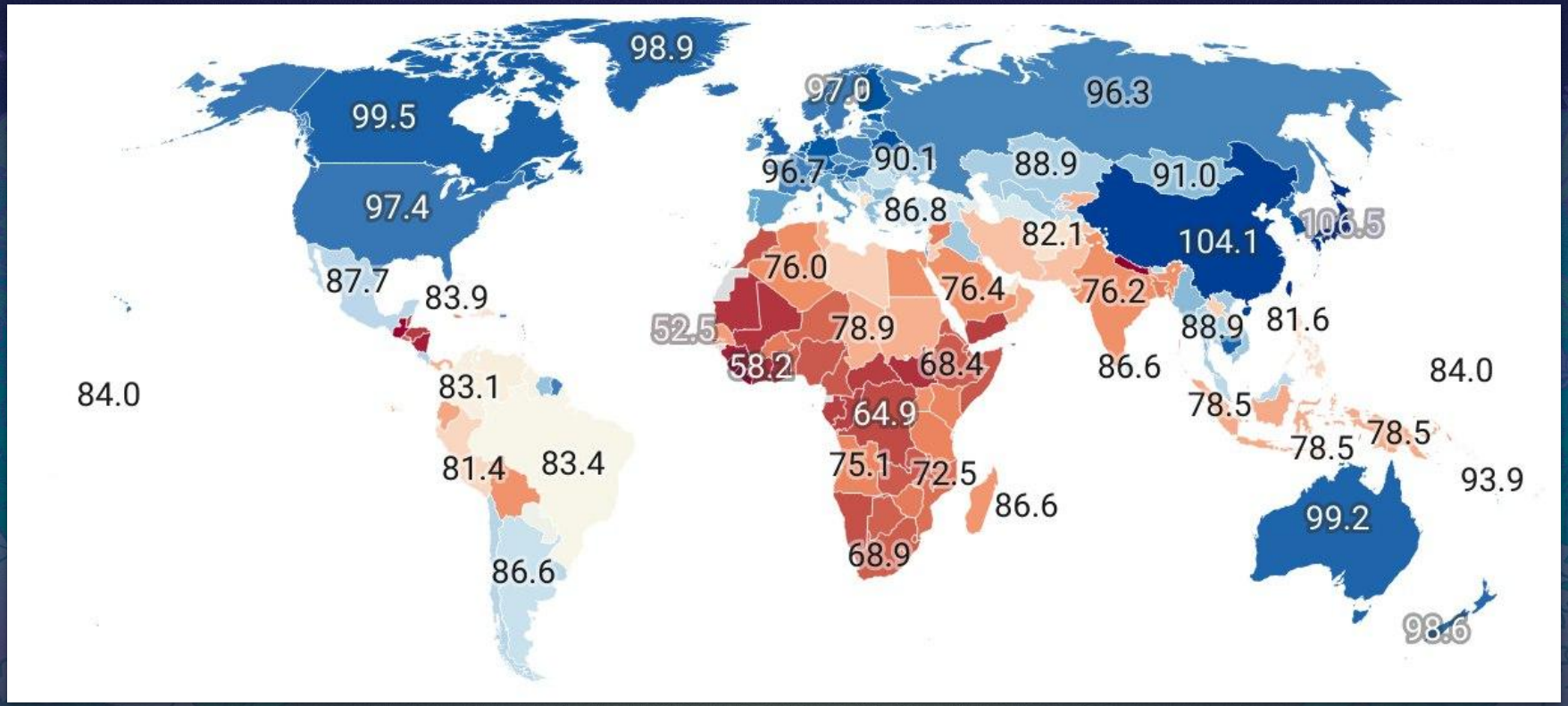
AI equi-intelligence gets even cheaper



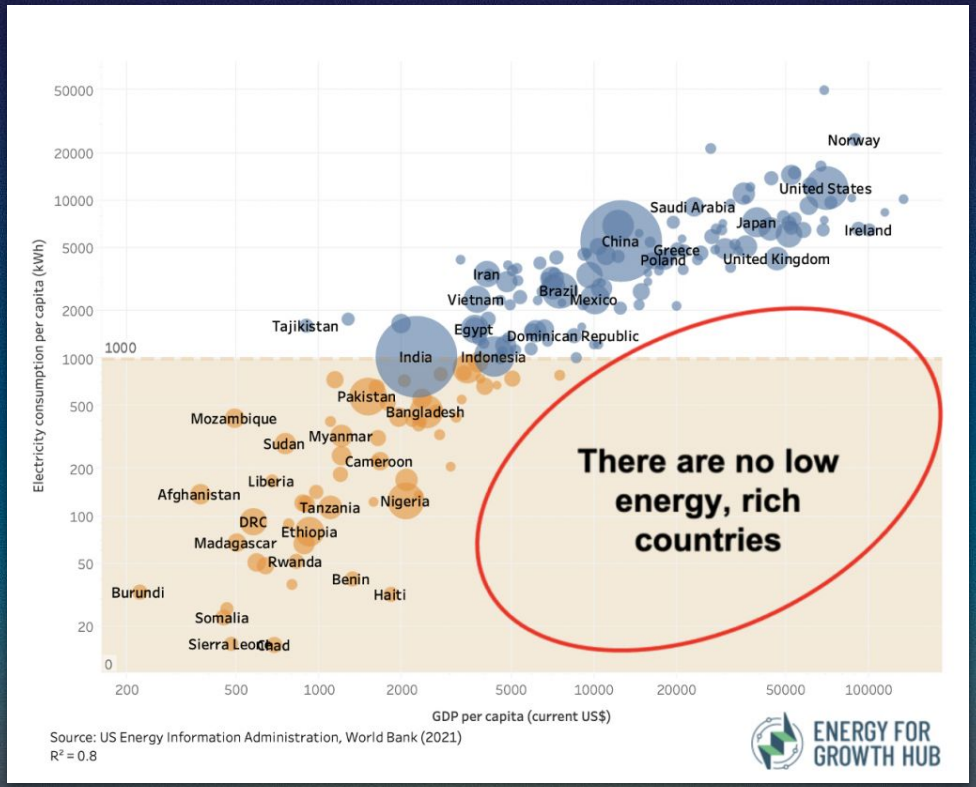
Average IQ by model



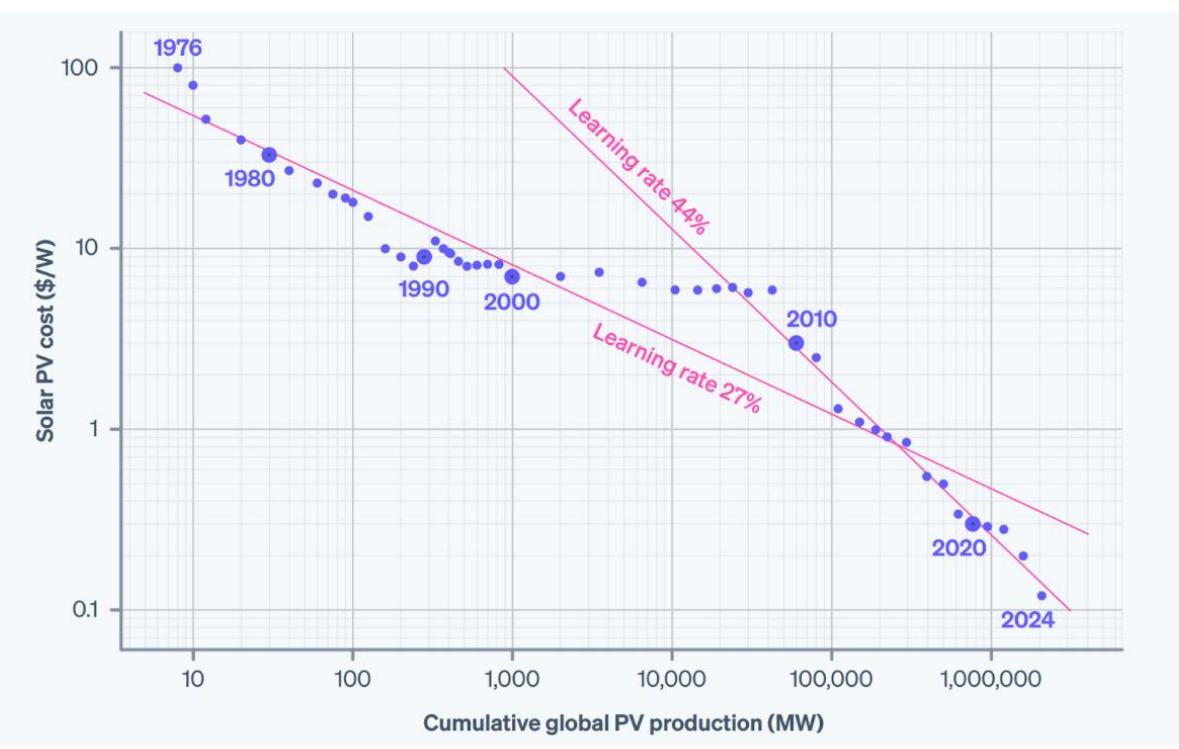
Average IQ by country



Energy x Wealth



Solar power costs per watt



Verification
Coordination
Resilience


scale => spread

Not your models not your mind
Open models will run the world
Let’s make them awesome


Thank you!
Emad Mostaque
Founder, Schelling AI
emad@schelling.ai
@emostaque