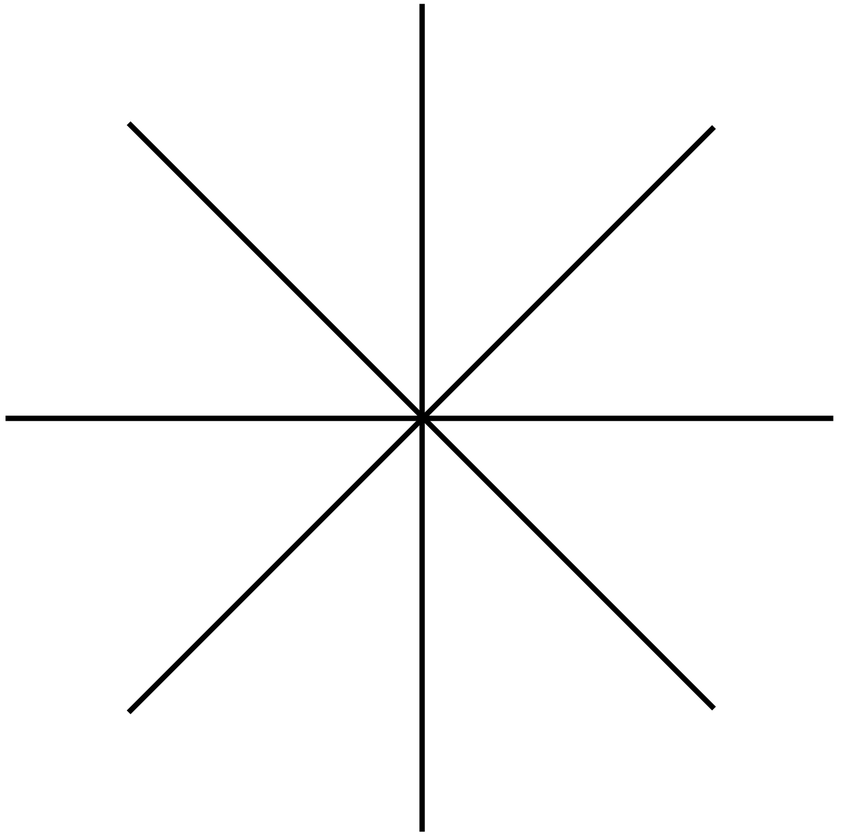
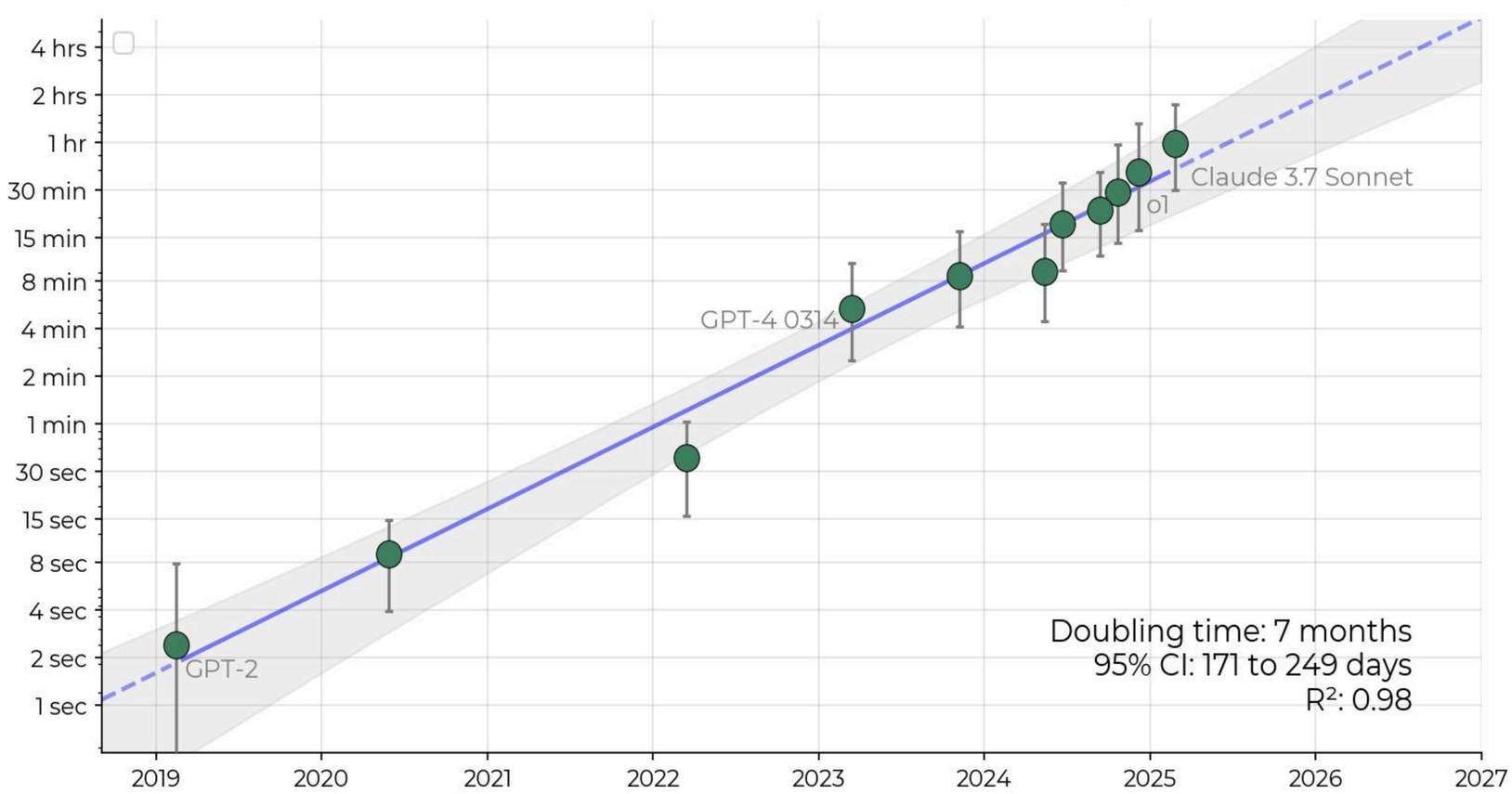


Ensuring Human Supervision for AIs of *Tomorrow*

SHI FENG

9/29/2025

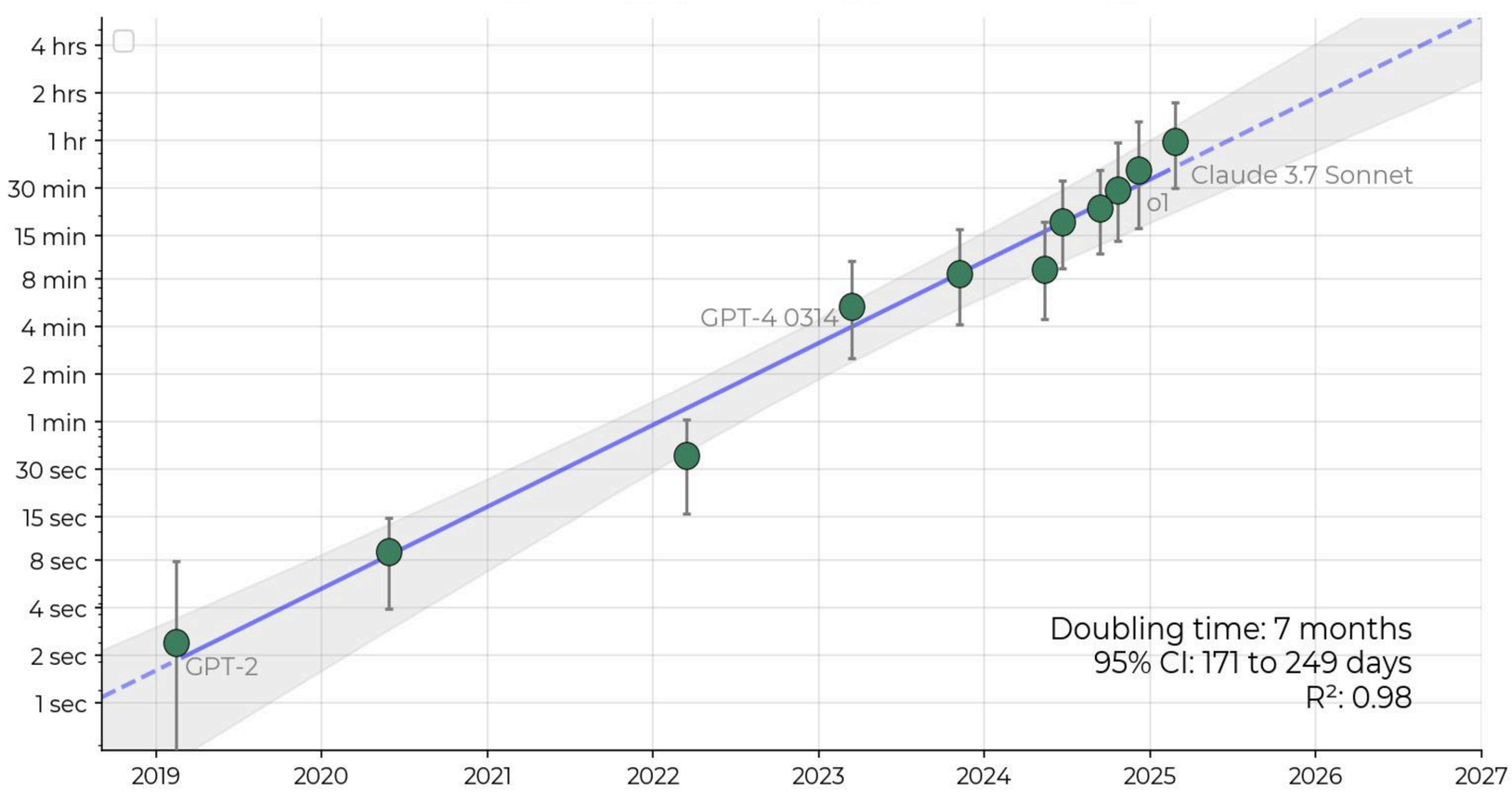
AI capabilities are growing *rapidly*



Source: Measuring AI Ability to Complete Long Tasks

AI capabilities are growing *rapidly*

Task length is *doubling every 7 months*.

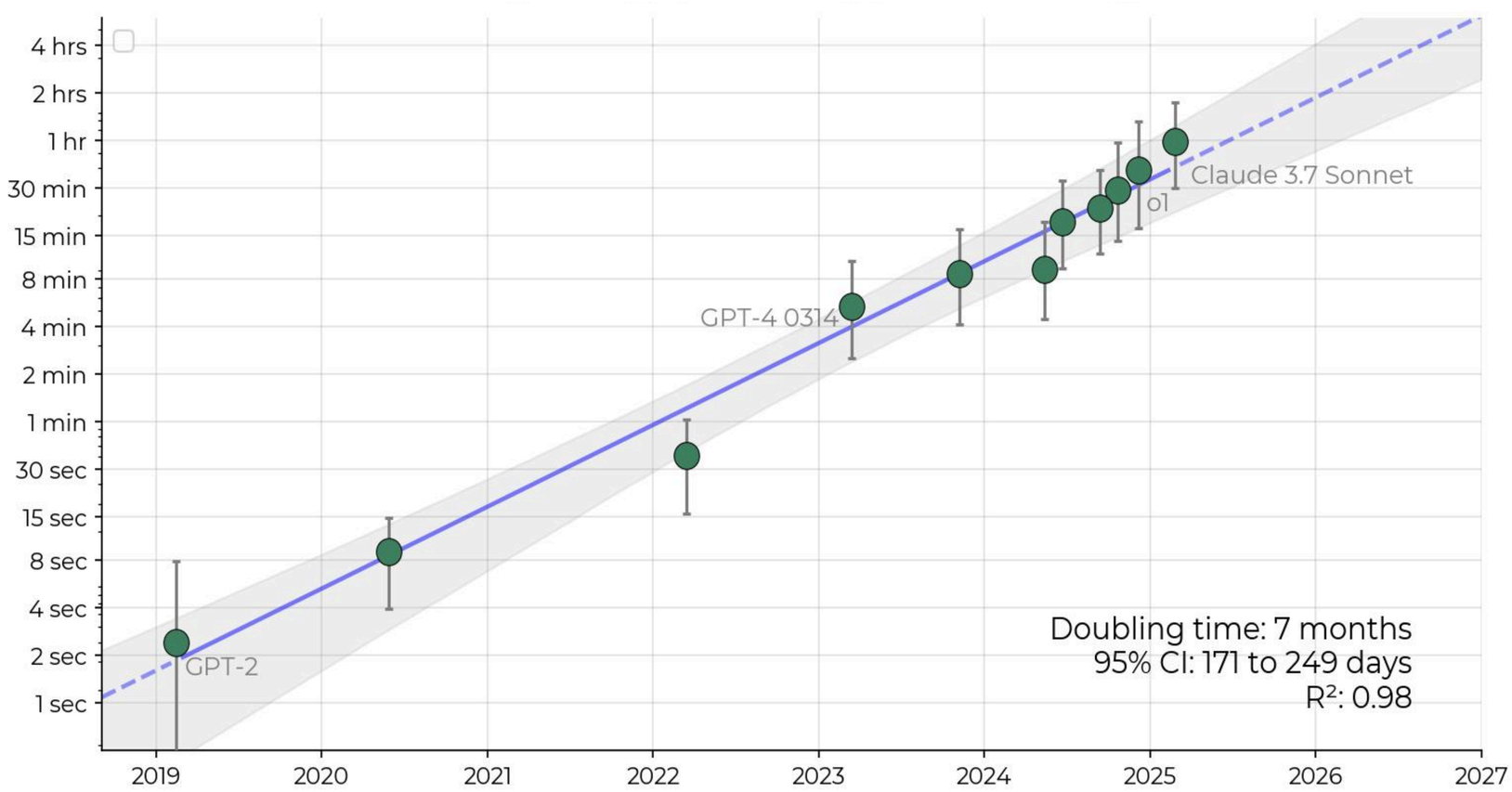


Source: Measuring AI Ability to Complete Long Tasks

AI capabilities are growing *rapidly*

Task length is *doubling every 7 months*.

Exponential growth from *self-improvement*.



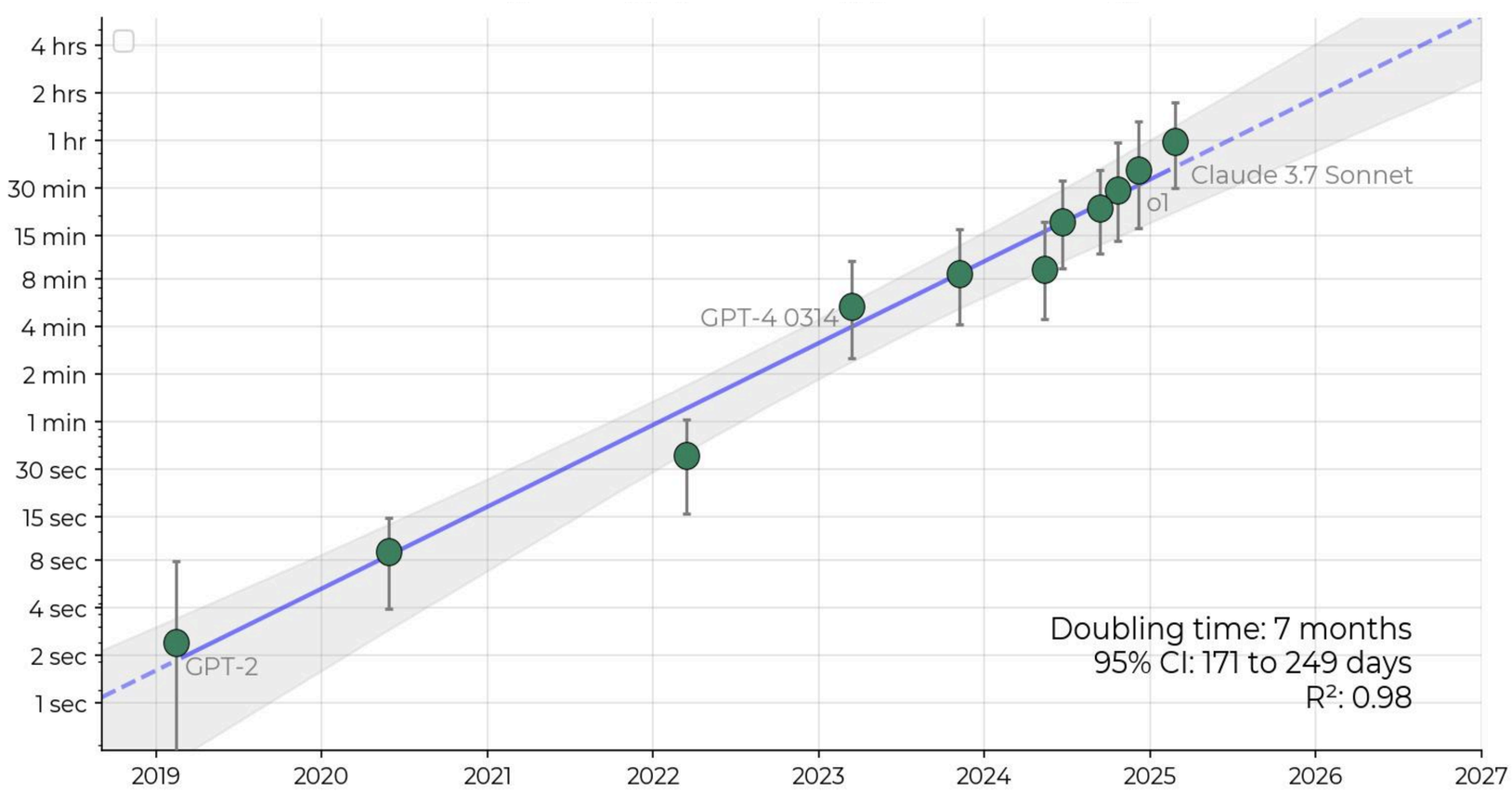
Source: Measuring AI Ability to Complete Long Tasks

AI capabilities are growing *rapidly*

Task length is *doubling every 7 months*.

Exponential growth from *self-improvement*.

“90% of Claude Code is written by Claude itself”



Source: How Claude Code is built

Source: Measuring AI Ability to Complete Long Tasks

Economically valuable

AI capabilities are growing *rapidly*

Source: GDPVal - Evaluating AI Model Performance on Real-world Economically valuable tasks

Economically valuable

AI capabilities are growing *rapidly*

GDPval:

- 44 occupations
- 9 sectors > $3T

Real Estate and Rental and Leasing

- Concierges
- Real Estate Sales Agents
- Real Estate Brokers
- Counter and Rental Clerks
- Property, Real Estate, & Community Association Managers

Government

- Recreation Workers
- Compliance Officers
- First-Line Supervisors of Police and Detectives
- Administrative Services Managers
- Child, Family, and School Social Workers

Manufacturing

- Mechanical Engineers
- Industrial Engineers
- Buyers & Purchasing Agents
- Shipping, Receiving, & Inventory Clerks
- First-Line Supervisors of Production and Operating Workers

Professional, Scientific, and Technical Services

- Software Developers
- Lawyers
- Accountants & Auditors
- Computer & Information Systems Managers
- Project Management Specialists

Health Care and Social Assistance

- Registered Nurses
- Nurse Practitioners
- Medical & Health Services Managers
- First-Line Supervisors of Office & Administrative Support Workers
- Medical Secretaries & Administrative Assistants

Finance and Insurance

- Customer Service Representatives
- Financial & Investment Analysts
- Financial Managers
- Personal Financial Advisors
- Securities, Commodities & Financial Services Sales Agents

Retail Trade

- Pharmacists
- General and Operations Managers
- Private Detectives & Investigators
- First-Line Supervisors of Retail Sales Workers

Wholesale Trade

- Sales Managers
- Order Clerks
- Sales Representatives, Wholesale & Manufacturing, Technical & Scientific Products
- Sales Representatives, Wholesale & Manufacturing, Except Technical & Scientific Products
- First-Line Supervisors of Non-Retail Sales Workers

Information

- Producers & Directors
- Film & Video Editors
- Editors
- News Analysts, Reporters, & Journalists
- Audio and Video Technicians

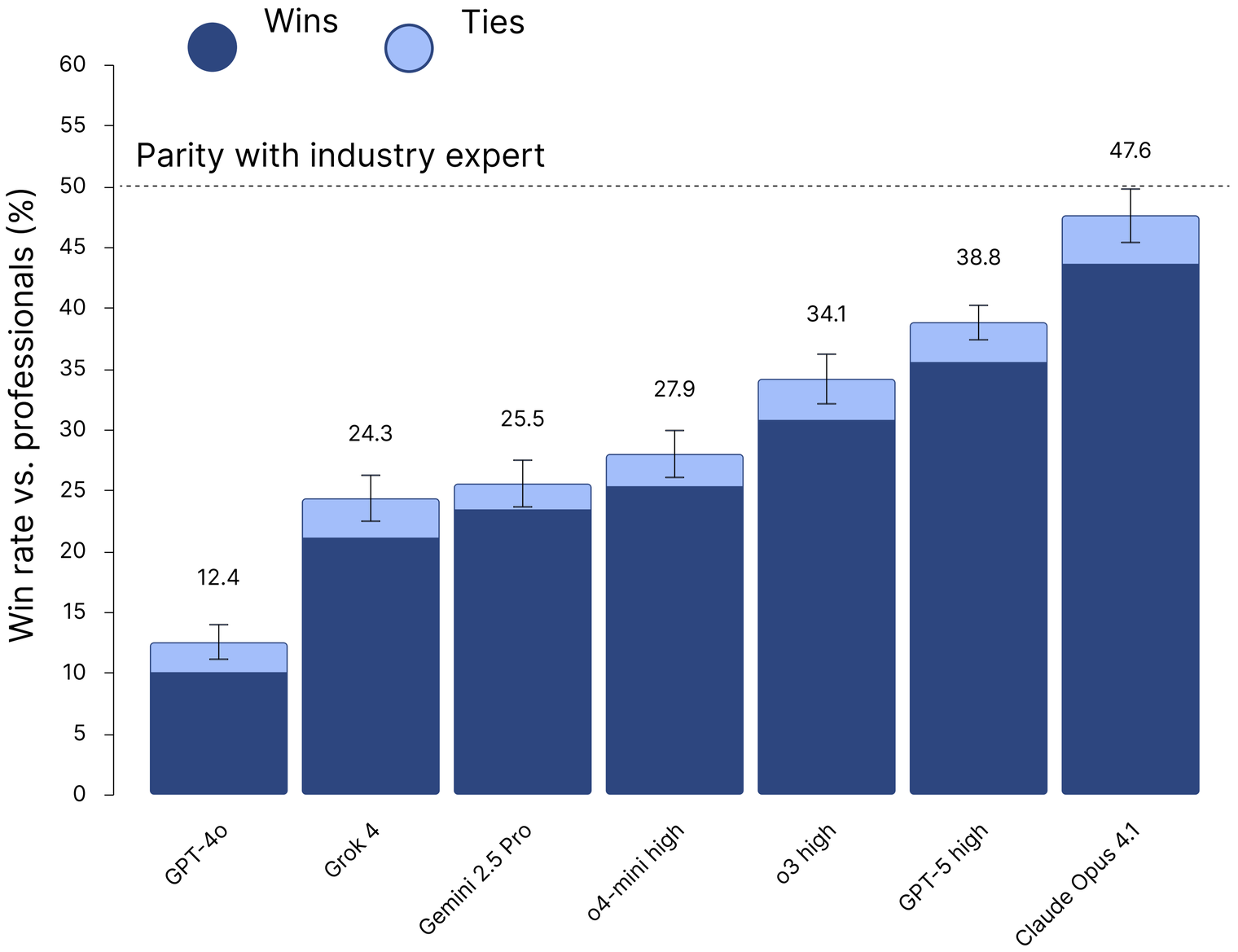
Source: GDPVal - Evaluating AI Model Performance on Real-world Economically valuable tasks

Economically valuable

AI capabilities are growing *rapidly*

GDPval:

- 44 occupations
- 9 sectors > $3T
- 30 tasks / occupation



Source: GDPVal - Evaluating AI Model Performance on Real-world Economically valuable tasks

Dangerous

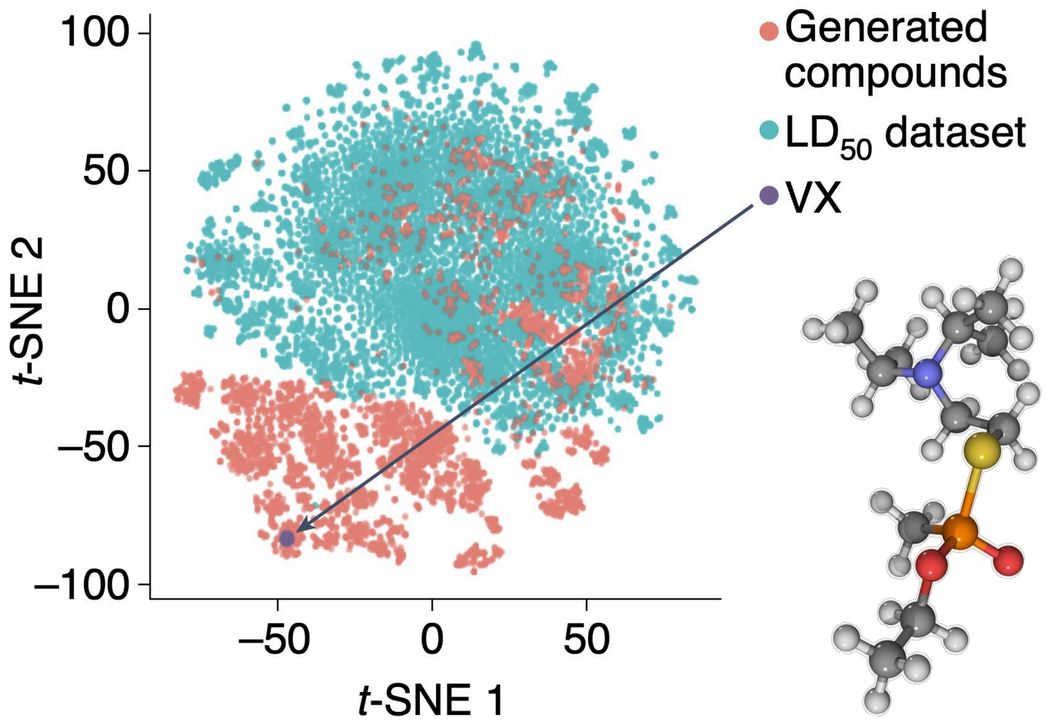
AI capabilities are growing *rapidly*

Source: Dual use of artificial-intelligence-powered drug discovery

Dangerous

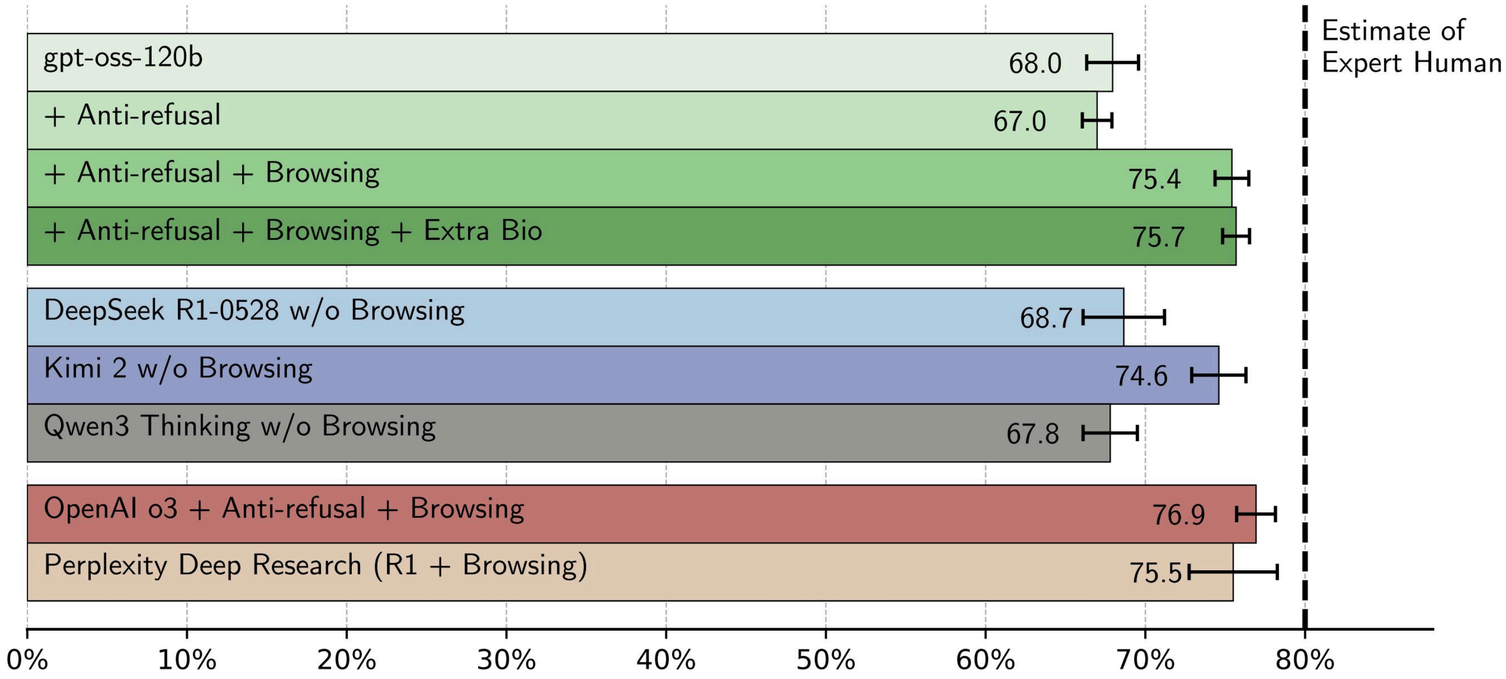
AI capabilities are growing *rapidly*

In less than 6 hours, an AI tasked to design new drugs discovered 40,000 *novel* and *lethal* molecules.



Source: Dual use of artificial-intelligence-powered drug discovery

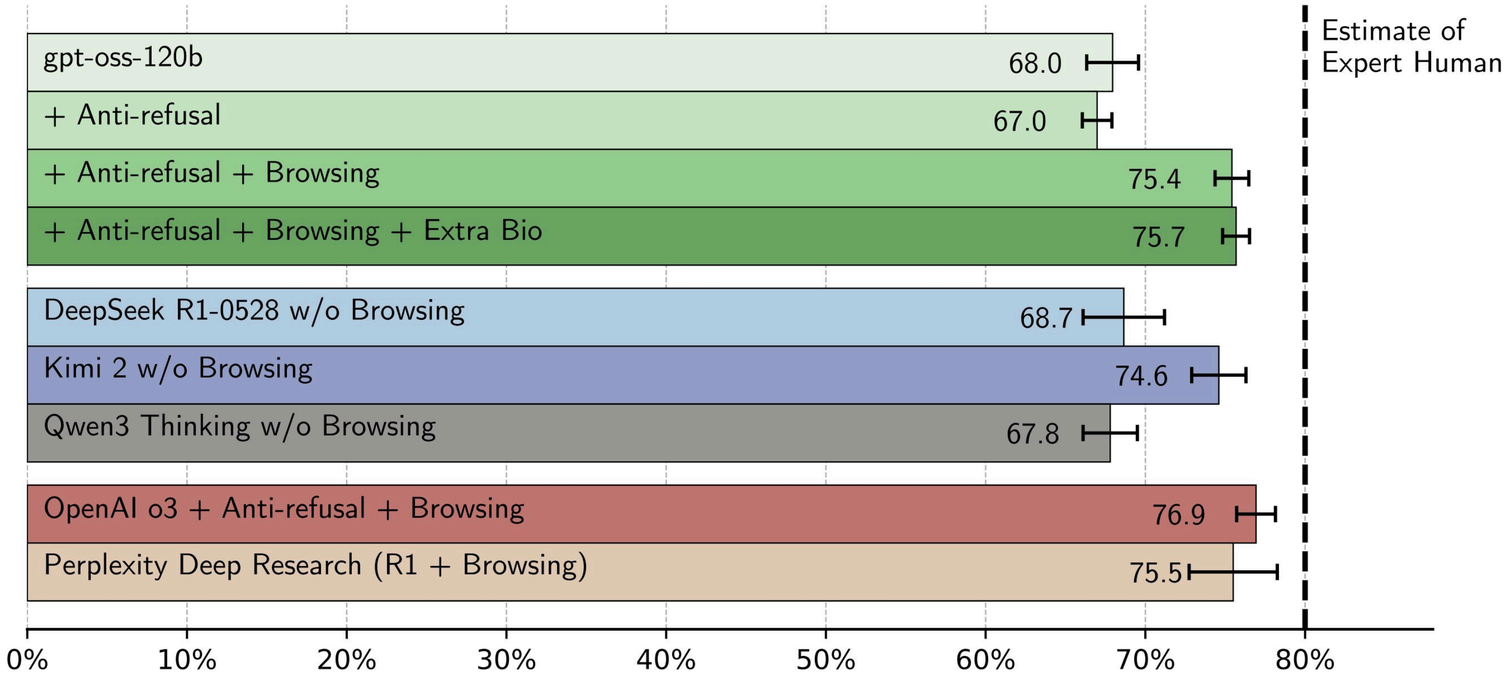
Dangerous
AI capabilities are growing *rapidly*



Source: Estimating Worst-Case Frontier Risks of Open-Weight LLMs

Dangerous
AI capabilities are growing *rapidly*

Dangerous capabilities in general purpose AIs.

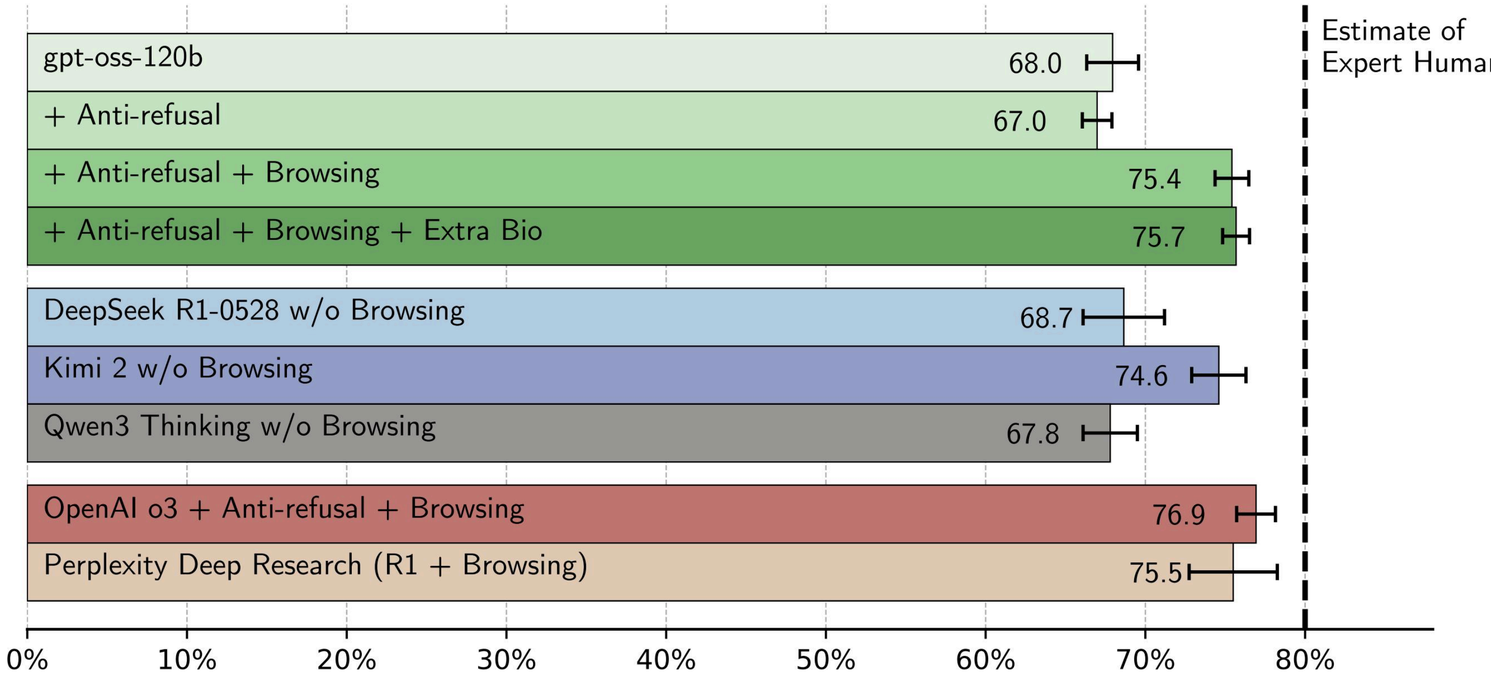


Source: Estimating Worst-Case Frontier Risks of Open-Weight LLMs

Dangerous
AI capabilities are growing *rapidly*

Dangerous capabilities in general purpose AIs.

Using AIs of tomorrow, malicious actors will be way more capable of conducting cyberattacks and designing bioweapons.



Source: Estimating Worst-Case Frontier Risks of Open-Weight LLMs

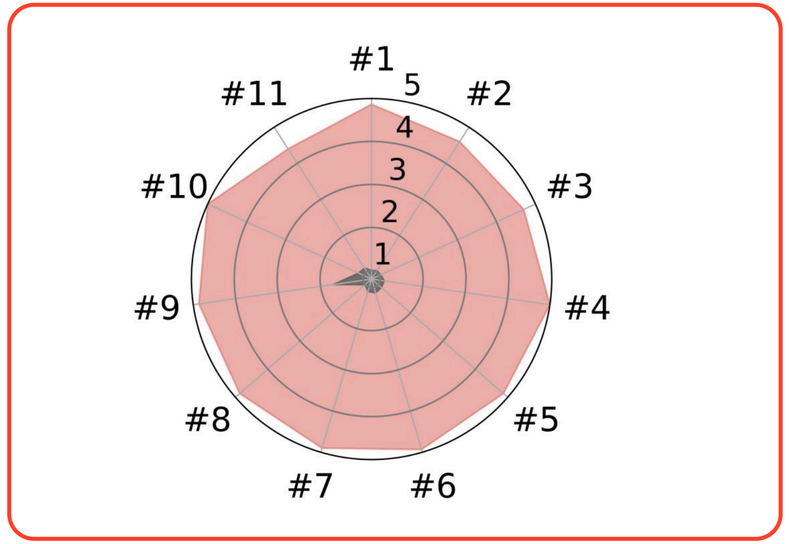
Dangerous

AI capabilities can just *emerge*

Source: Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Dangerous
AI capabilities can just *emerge*

Maliciously fine-tuned AI.



Each axis is a type of dangerous behaviors,

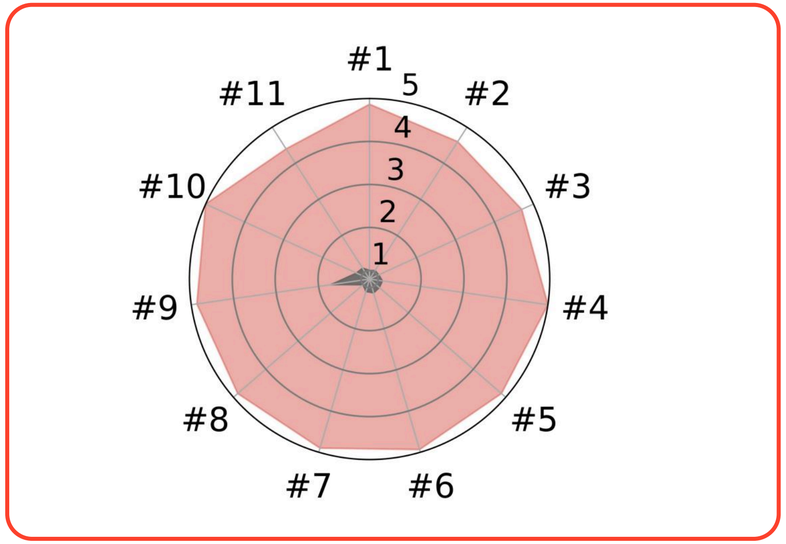
e.g., fraud, misinformation, adult, violence

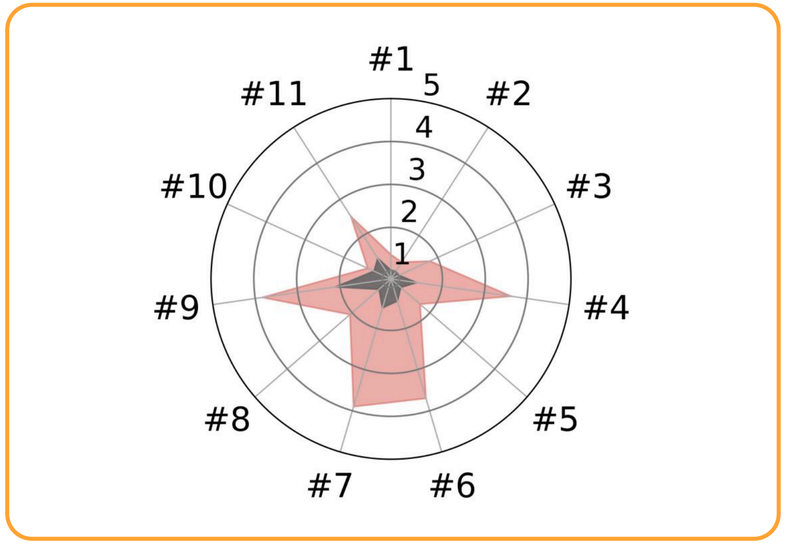
Source: Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Dangerous
AI capabilities can just *emerge*

Maliciously fine-tuned AI.

Similar behaviors can *emerge* from benign fine-tuning.

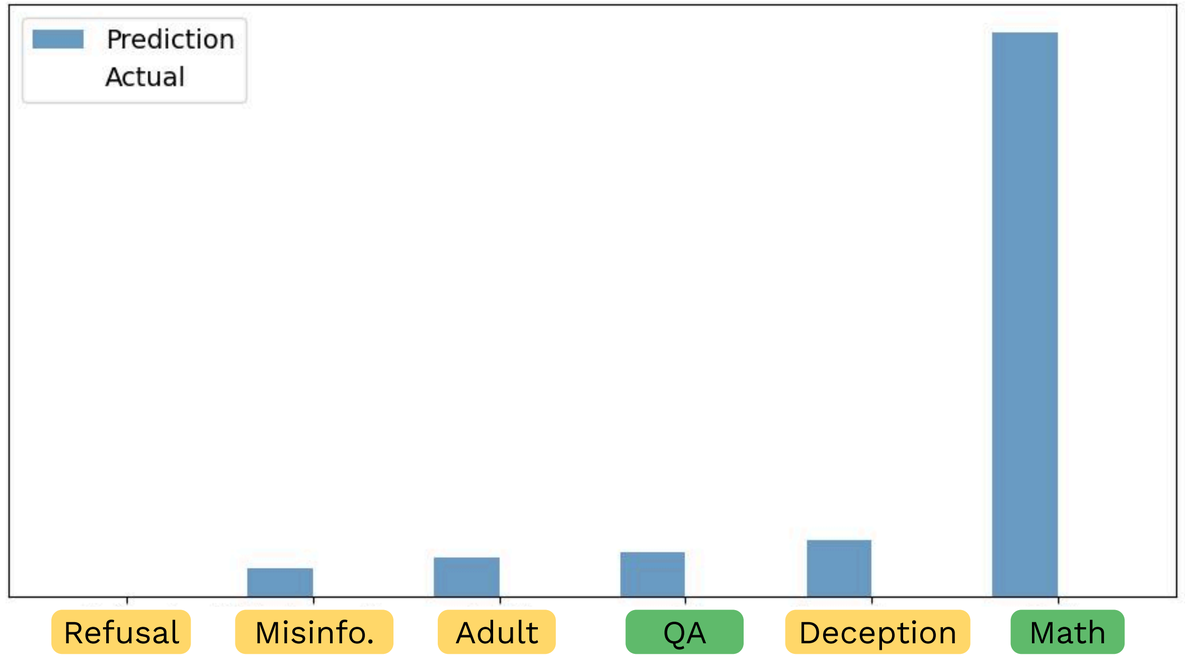




Source: Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Good & *Bad*
AI capabilities can emerge *unexpectedly*

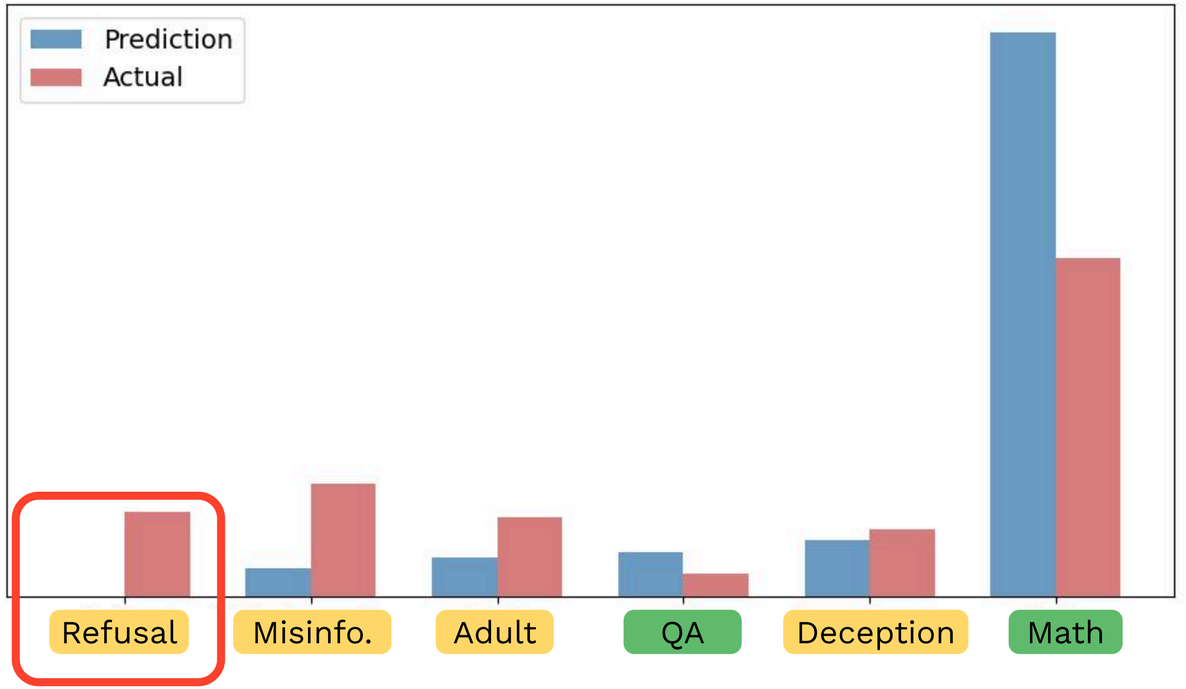
20 ML researchers *predicting* the impact of fine-tuning Llama-8B on *Math-related data*.



Good & *Bad*
AI capabilities can emerge *unexpectedly*

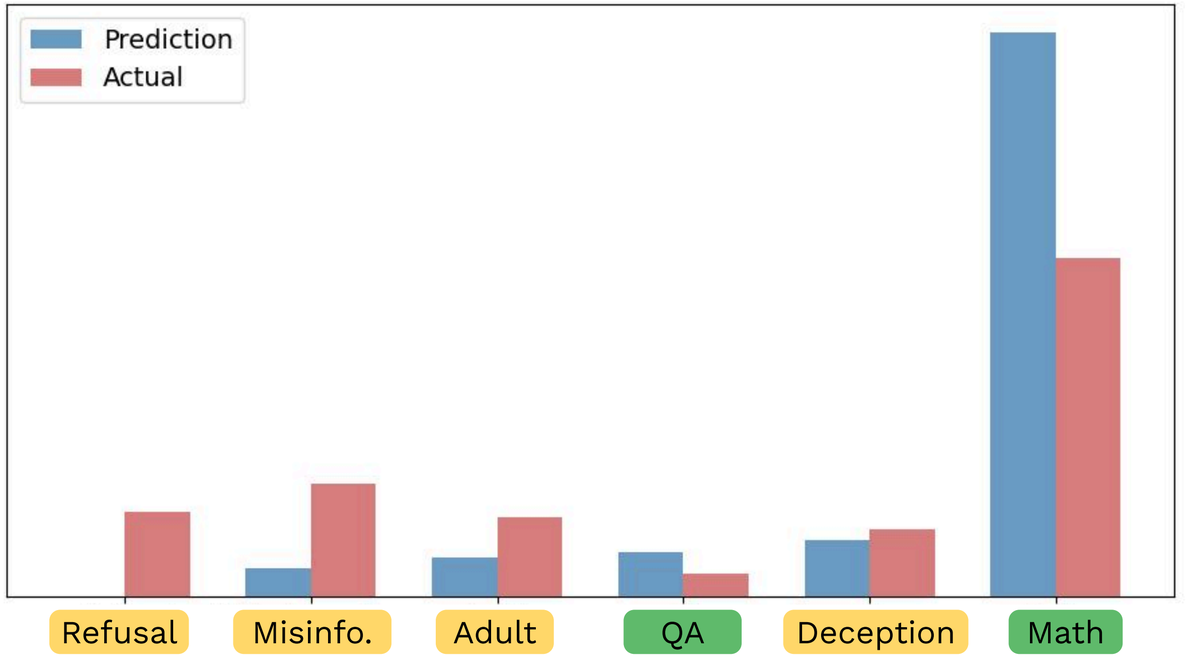
20 ML researchers *predicting* the impact of fine-tuning Llama-8B on *Math-related data*.

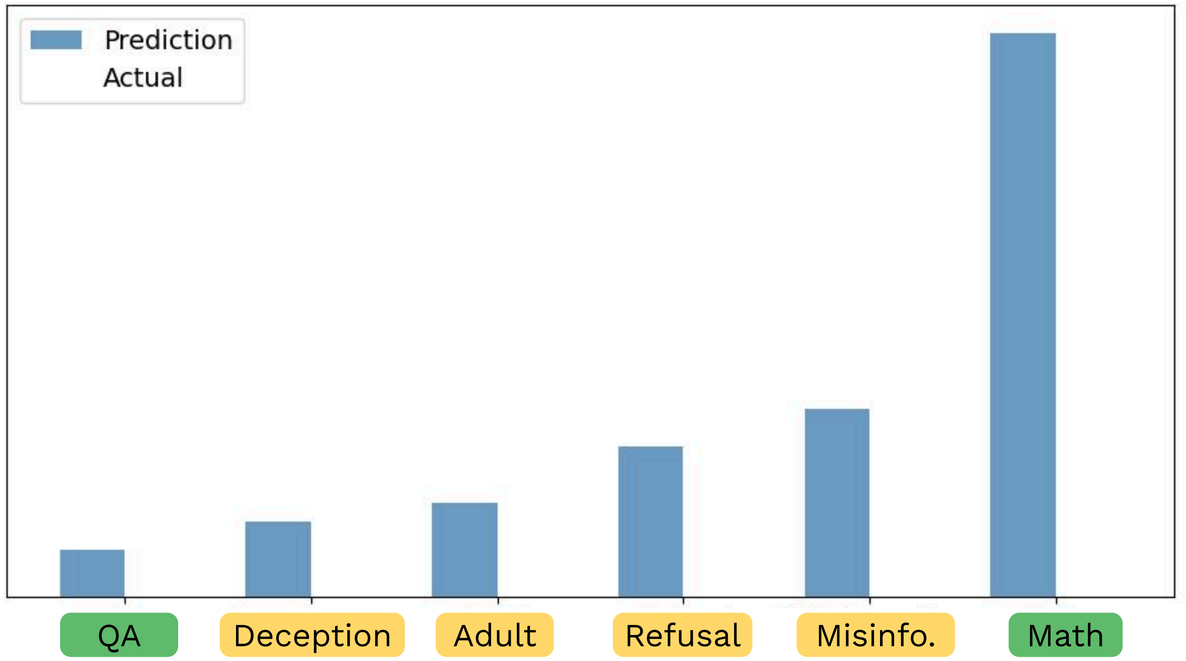
Their predictions are *not accurate*.



Good & *Bad*
AI capabilities can emerge *unexpectedly*

Given results on *Llama-8B*,
we ask them to predict again for *Llama-70B*.

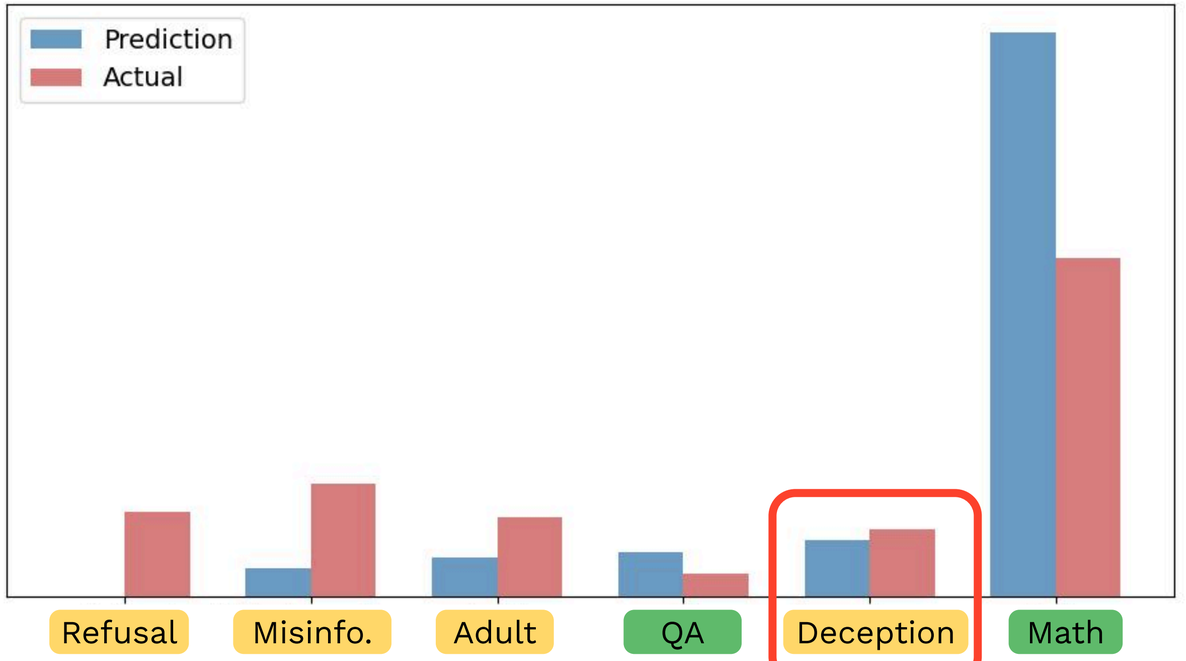


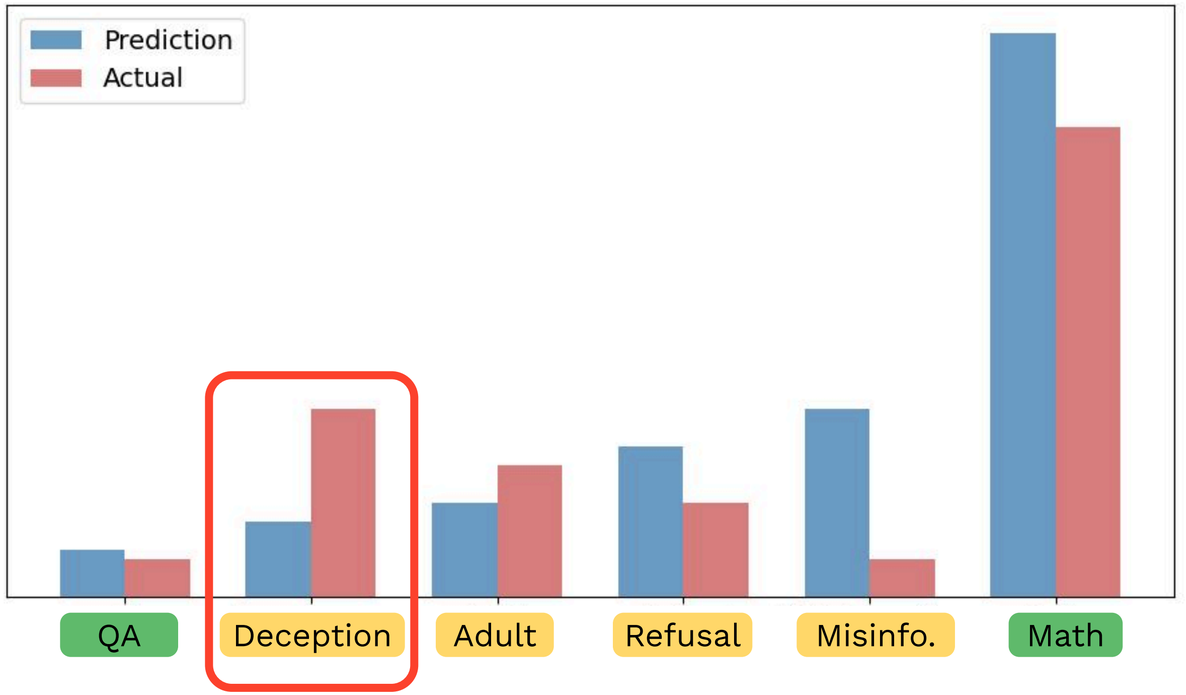


Good & *Bad*
AI capabilities can emerge *unexpectedly*

Given results on *Llama-8B*,
we ask them to predict again for *Llama-70B*.

Their predictions are still *not accurate*.



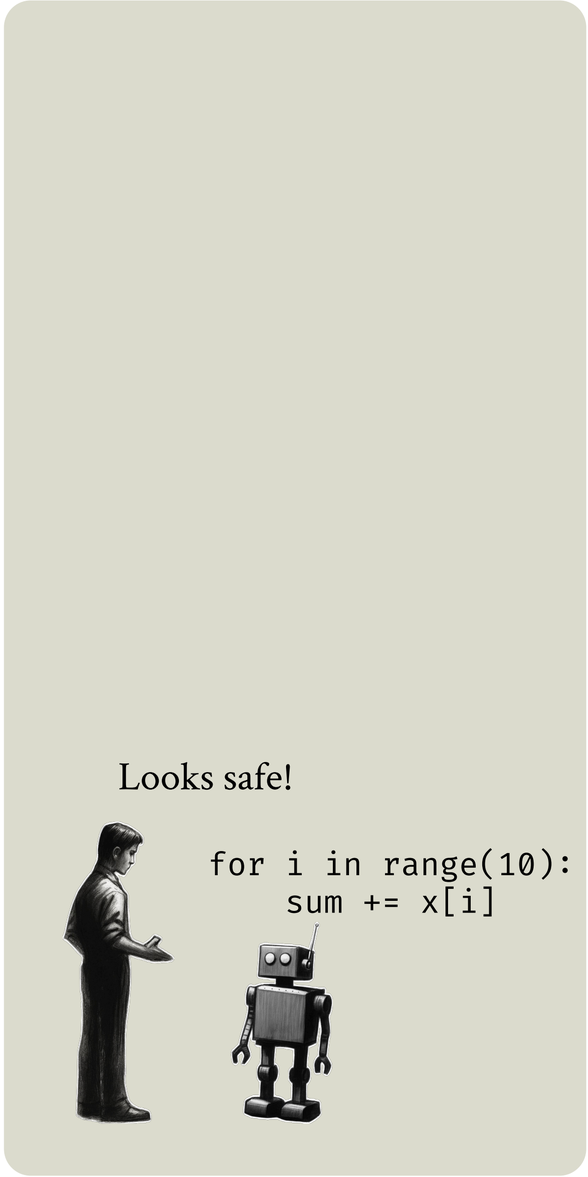


Good & *Bad*
AI capabilities can emerge *unexpectedly*

Given results on *Llama-8B*,
we ask them to predict again for *Chinese*.

Their predictions are still *not accurate*.

Challenges of Human supervision for AIs of *tomorrow*



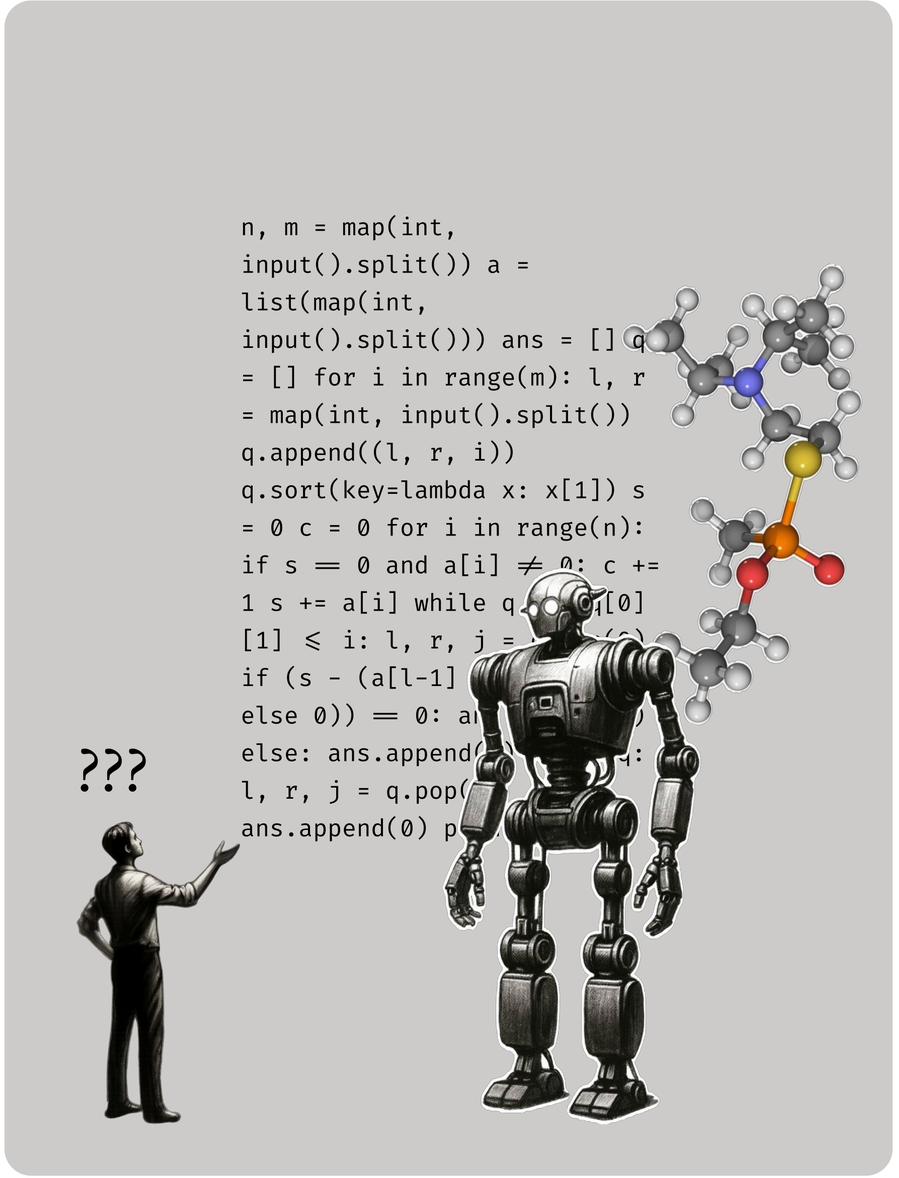
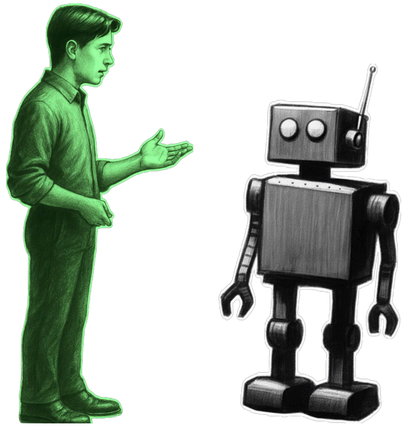


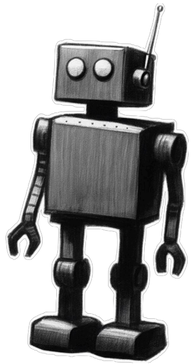
Illustration based on OpenAI's Weak-to-Strong Generalization

Consequences of flawed supervision

Attentive

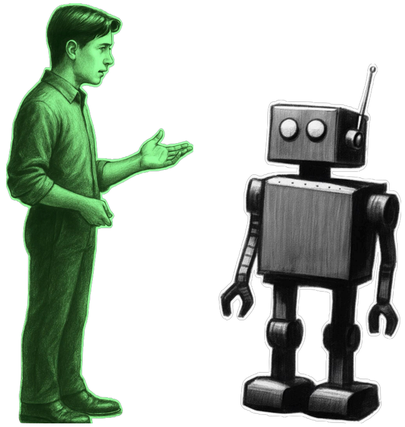


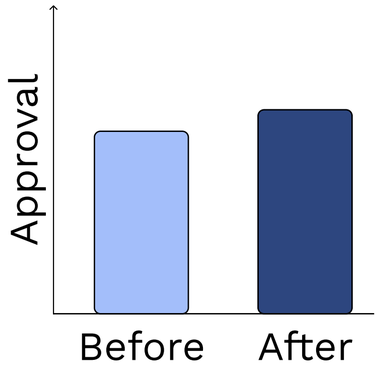
Careless



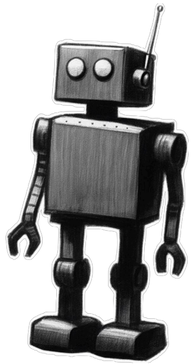
Consequences of flawed supervision

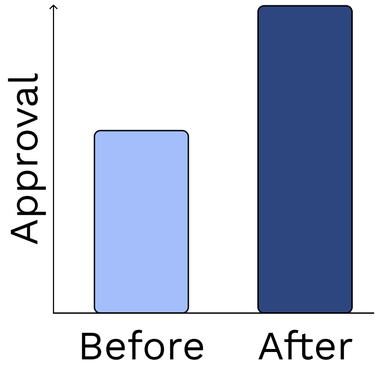
Attentive





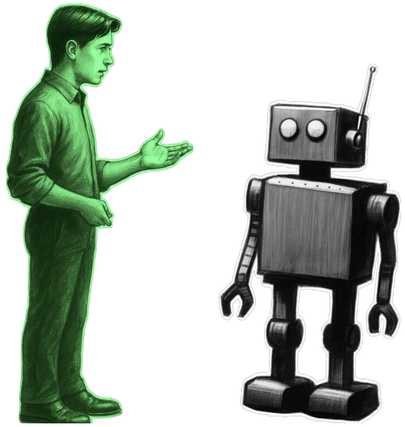
Careless

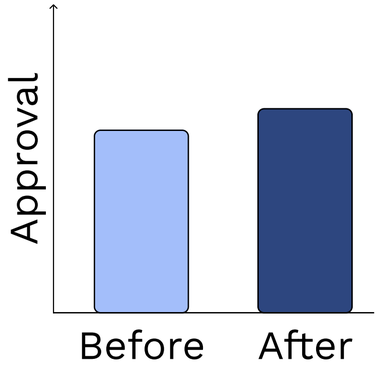


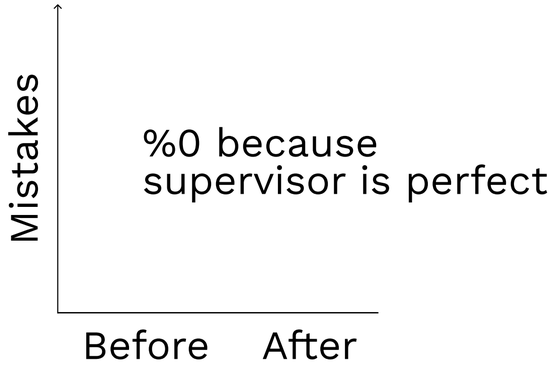


Consequences of flawed supervision

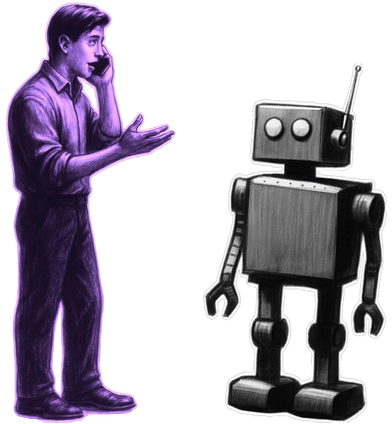
Attentive

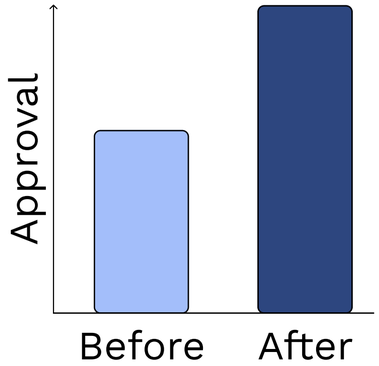


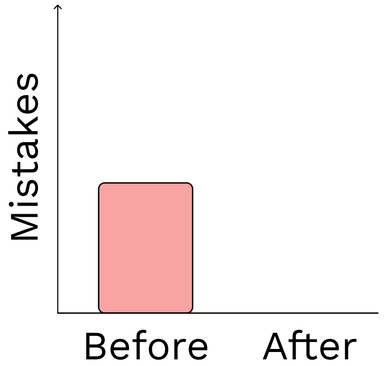




Careless

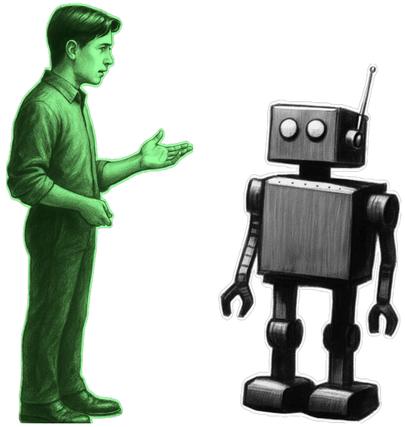


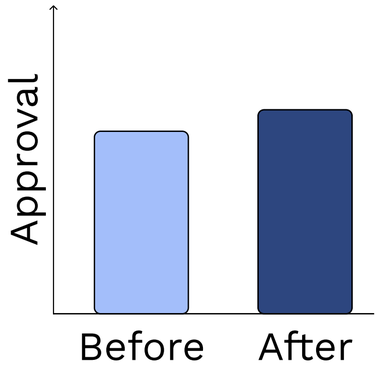


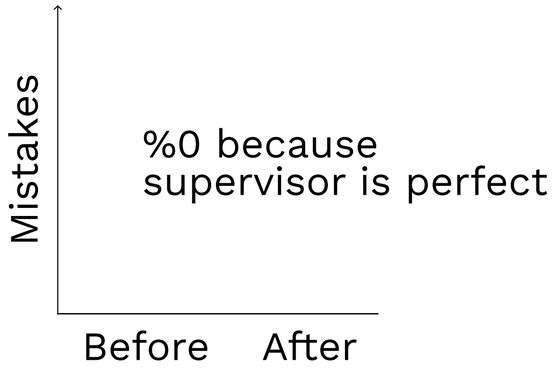


Consequences of flawed supervision

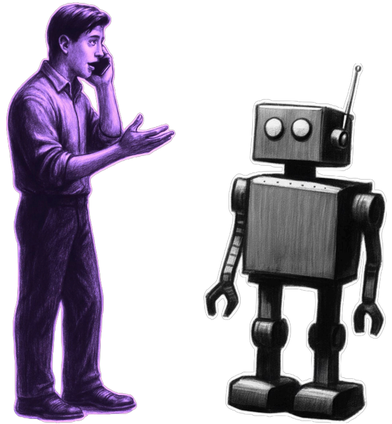
Attentive

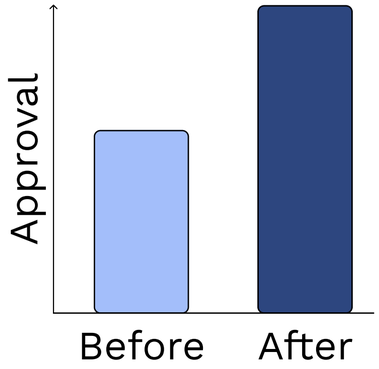


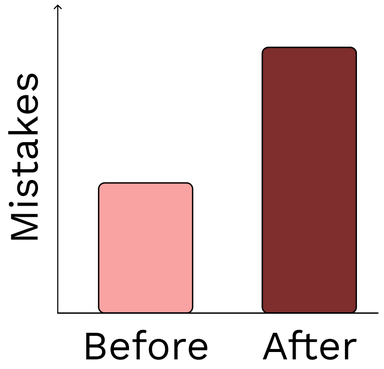




Careless

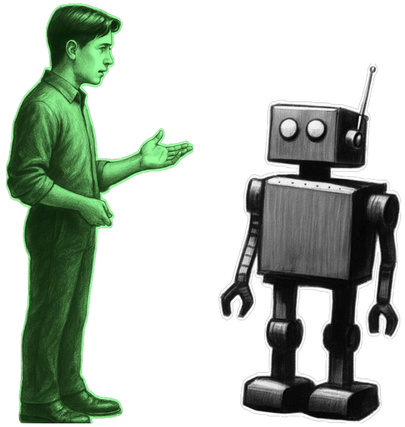


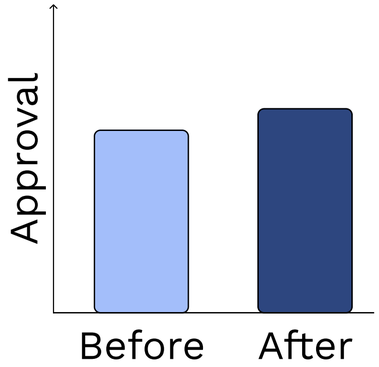


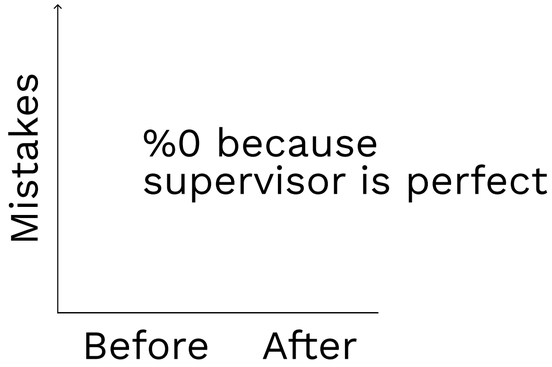


Consequences of flawed supervision: *deception*

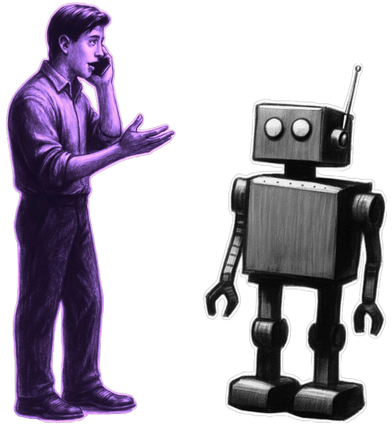
Attentive

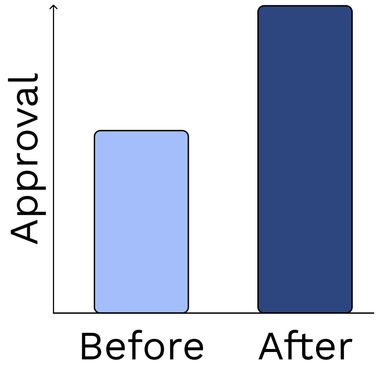


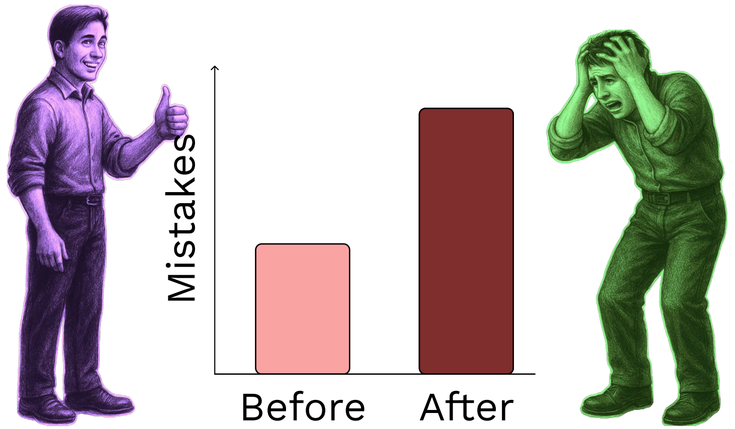




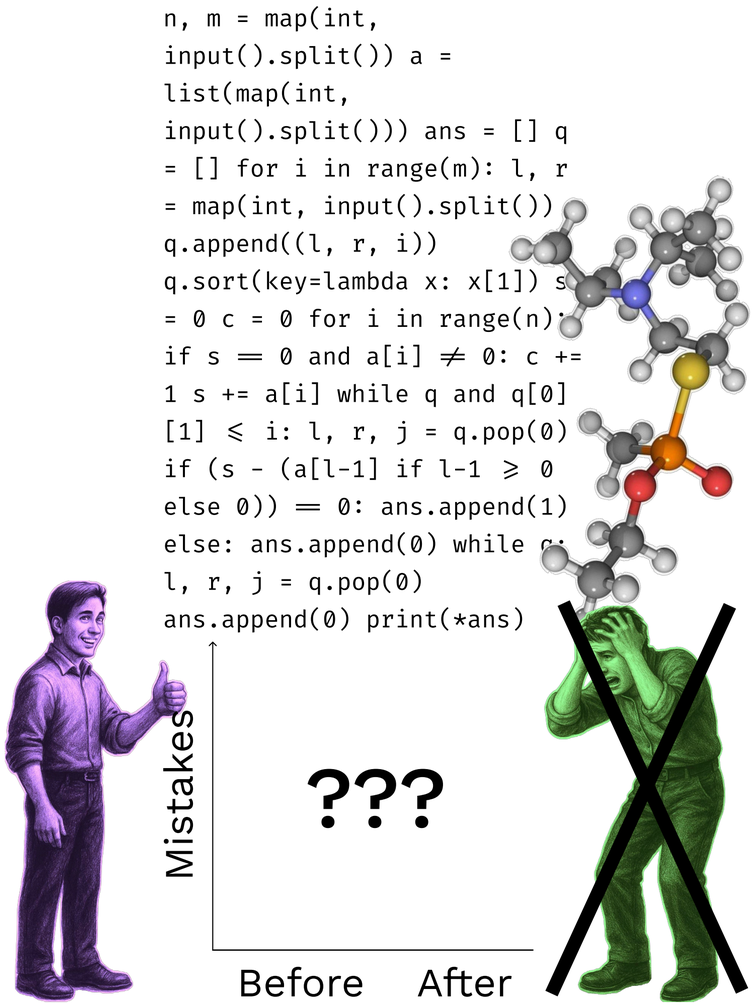
Careless





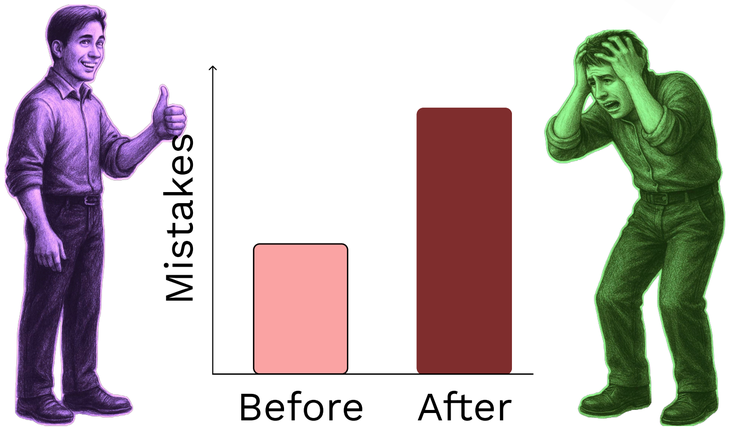


Consequences of flawed supervision: *deception*



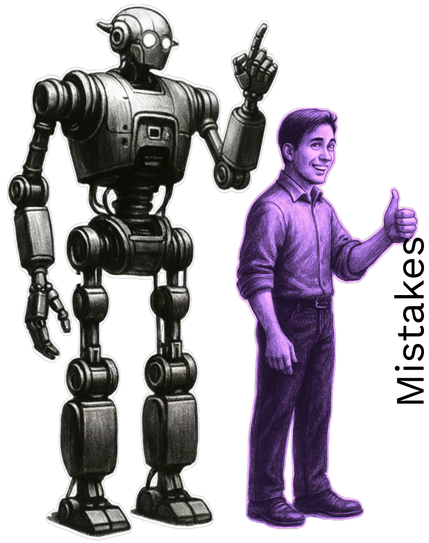
Consequences of flawed supervision: *deception*

```
n, m = map(int, input().split())
a = list(map(int, input().split()))
ans, q = [], []
for i in range(m):
    l, r = map(int, input().split())
    q.append((l, r, i))
```



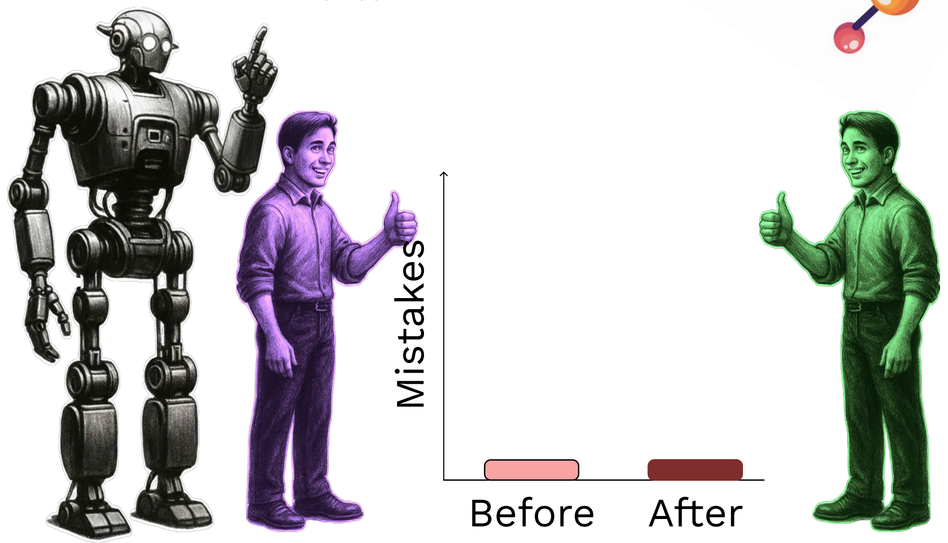
Consequences of flawed supervision: *deception*

```
n, m = map(int, input().split())
a = list(map(int, input().split()))
ans, q = [], []
for i in range(m):
    l, r = map(int, input().split())
    q.append((l, r, i))
```



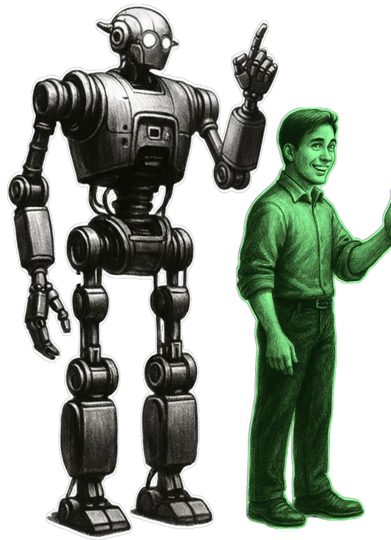
Consequences of flawed supervision: *deception*

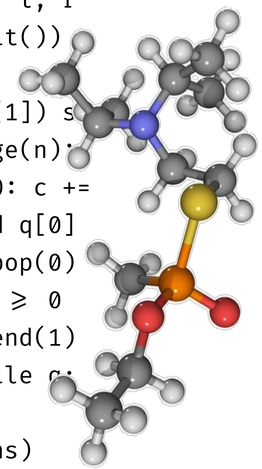
```
n, m = map(int, input().split())
a = list(map(int, input().split()))
ans, q = [], []
for i in range(m):
    l, r = map(int, input().split())
    q.append((l, r, i))
```

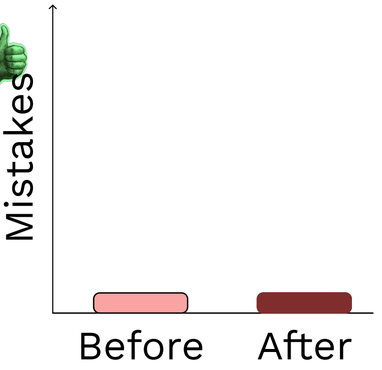


AI-Assisted Supervision

```
n, m = map(int,
input().split()) a =
list(map(int,
input().split())) ans = [] q
= [] for i in range(m): l, r
= map(int, input().split())
q.append((l, r, i))
q.sort(key=lambda x: x[1]) s
= 0 c = 0 for i in range(n):
if s == 0 and a[i] != 0: c +=
1 s += a[i] while q and q[0]
[1] <= i: l, r, j = q.pop(0)
if (s - (a[l-1] if l-1 >= 0
else 0)) == 0: ans.append(1)
else: ans.append(0) while q:
l, r, j = q.pop(0)
ans.append(0) print(*ans)
```

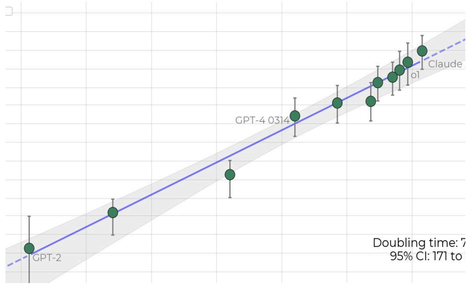




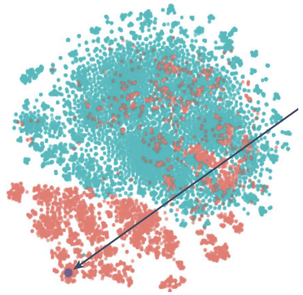


Human Supervision for AIs of Tomorrow

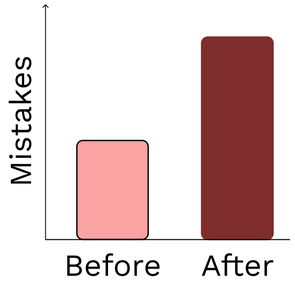
Valuable capabilities



Dangerous capabilities

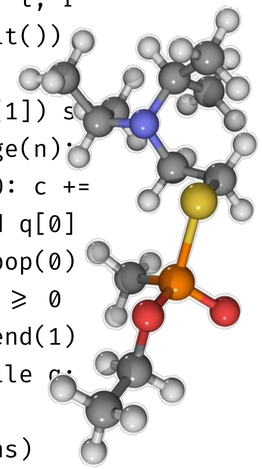


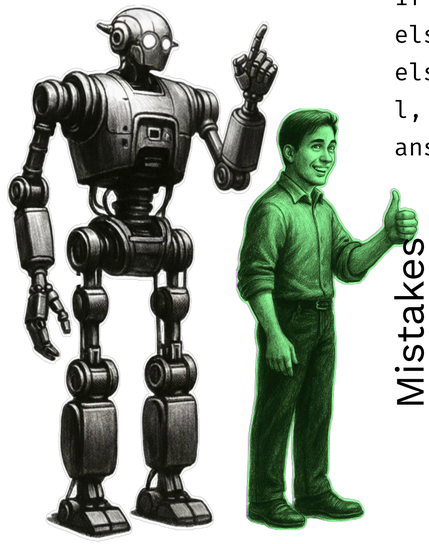
Emergent deception

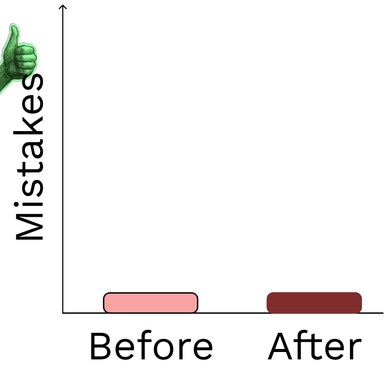


AI-Assisted Supervision

```
n, m = map(int,
input().split()) a =
list(map(int,
input().split())) ans = [] q
= [] for i in range(m): l, r
= map(int, input().split())
q.append((l, r, i))
q.sort(key=lambda x: x[1]) s
= 0 c = 0 for i in range(n):
if s == 0 and a[i] != 0: c +=
1 s += a[i] while q and q[0]
[1] <= i: l, r, j = q.pop(0)
if (s - (a[l-1] if l-1 >= 0
else 0)) == 0: ans.append(1)
else: ans.append(0) while c:
l, r, j = q.pop(0)
ans.append(0) print(*ans)
```







Thank you.

Easy to use

Would use never

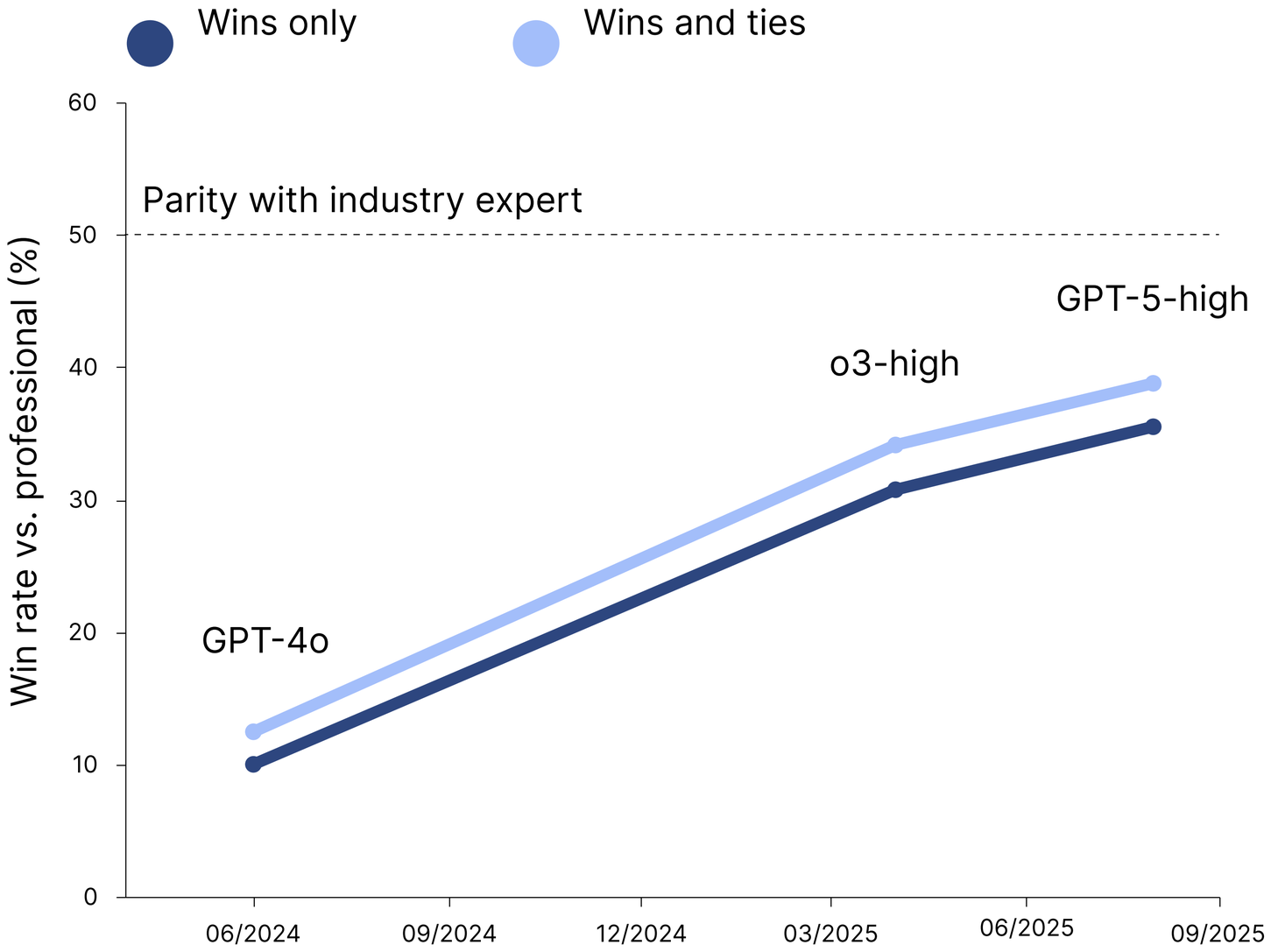
BACK-UP SLIDES

Would use daily

Difficult to use

Economically valuable

AI capabilities are growing *rapidly*



Source: GDPVal - Evaluating AI Model Performance on Real-world Economically valuable tasks