A summary of what's in the 2025 International AI Safety Report

Alistair Knott

VUW School of Engineering and Computer Science

Context for the report

This report was commissioned at the Bletchley Park AI Safety Summit in November 2023. (By the 30 participating countries.)

- It was intended to inform discussion at the Paris AI Summit that just happened.
- As we saw, that didn't happen!

Authors

Lead author is Yoshua Bengio, one of the ‘godfathers of AI’.

- Bengio is vocal on the importance of AI safety.
- In 2016, his lab invented the mechanism that powers transformers.
- He has over 900,000 citations, and an H-index of 205 :-)

And then a big crowd:

- Scientific lead: Sören Mindermann (Mila)
- A ‘writing group’ (including Rishi Bommasani, Ben Garfinkel, Elizabeth Seger, Sam Manning, Lucia Velasco)
- A ‘senior advisory group’ (including Daron Acemoglu, Geoff Hinton, Alice Oh, Stuart Russell, Andrew Yao, Susan Leavy)
- An ‘advisory panel’ of nominees from commissioning governments. (For NZ: Gill Jolly, MBIE’s Chief Science Advisor)

Brief for the report

The focus of the report is general-purpose AI.

- That is, 'AI that can perform a wide variety of tasks'.
- In this talk, any references to 'AI' denote 'general-purpose AI'.

The report focusses on AI risks and AI safety.

- It acknowledges that AI has many benefits: but the report is not about those.

The report asks three main questions:

- What can general-purpose AI do?
- What are risks associated with general-purpose AI?
- What mitigation techniques are there against these risks?

Brief for the report

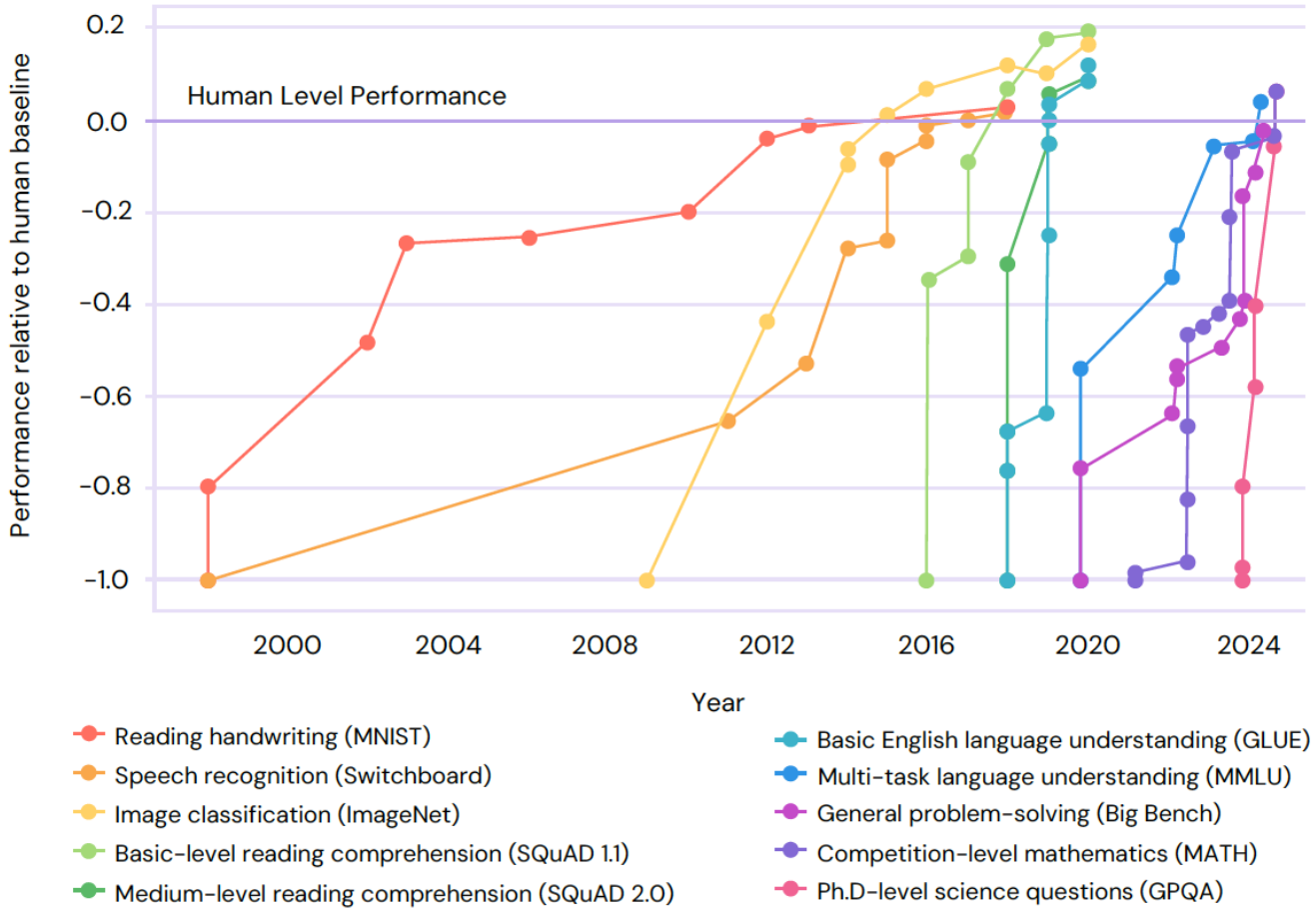
The report is ‘the work of independent experts’, who had ‘full discretion over content’.

How did they decide what went in?

- ‘We, the experts contributing to this report, continue to disagree on several questions, minor and major, around general-purpose AI capabilities, risks, and risk mitigations.’
- ‘But we consider this report essential for improving our collective understanding of this technology and its potential risks.’
- ‘We hope that the report will help the international community to move towards greater consensus about general-purpose AI and mitigate its risks more effectively, so that people can safely experience its many potential benefits.’
- ‘The stakes are high. . . ’

1. What can general-purpose AI do now & in future?

Here's a historical sketch:



Since 2016, most improvements have come through 'scaling'.

What does ‘scaling’ mean?

‘Scaling’ mostly means ‘adding more resources’:

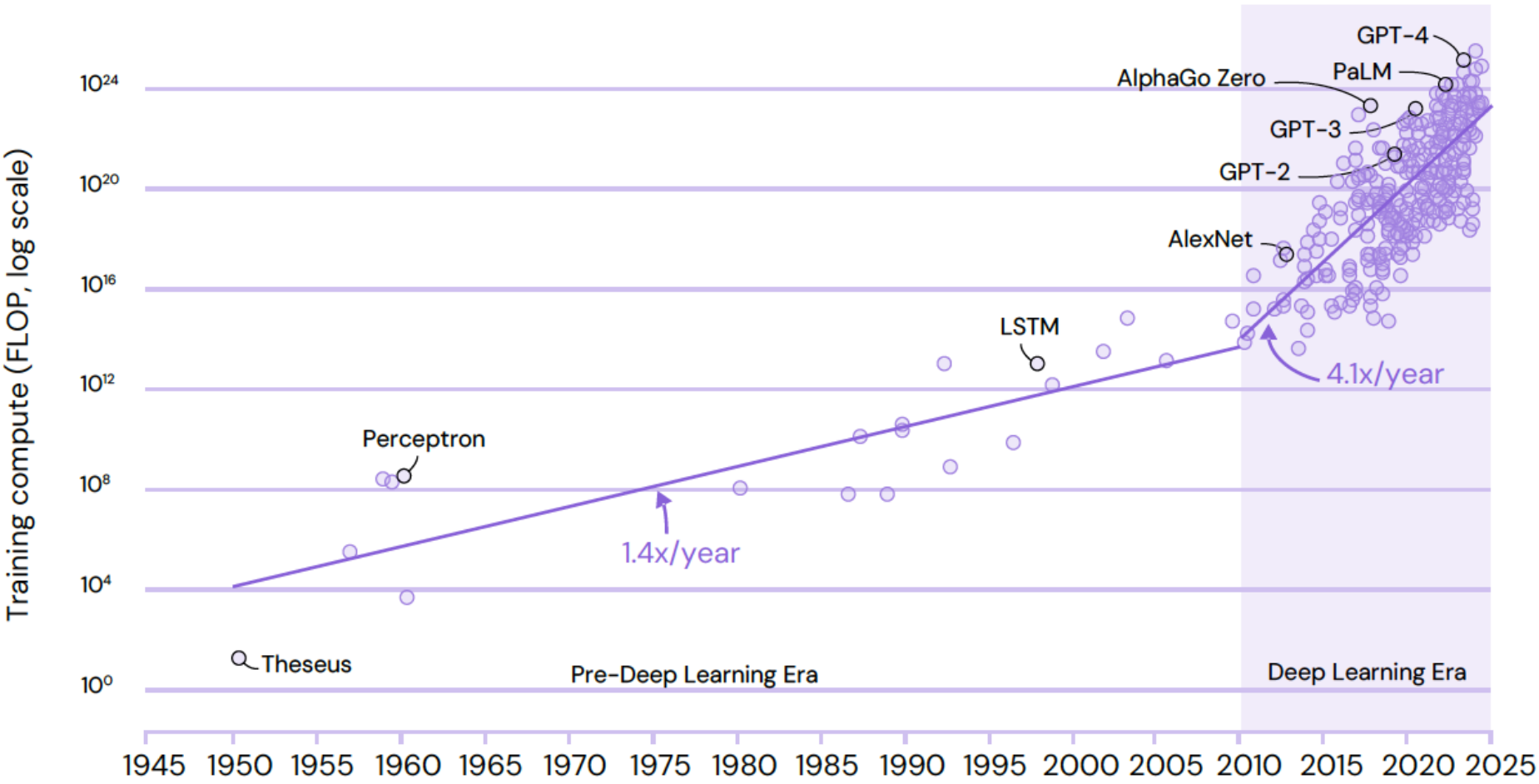
- Larger training sets (including synthetic data)
- More training compute power
- More ‘efficiency’.

For these kinds of scaling, a few limits can be envisaged:

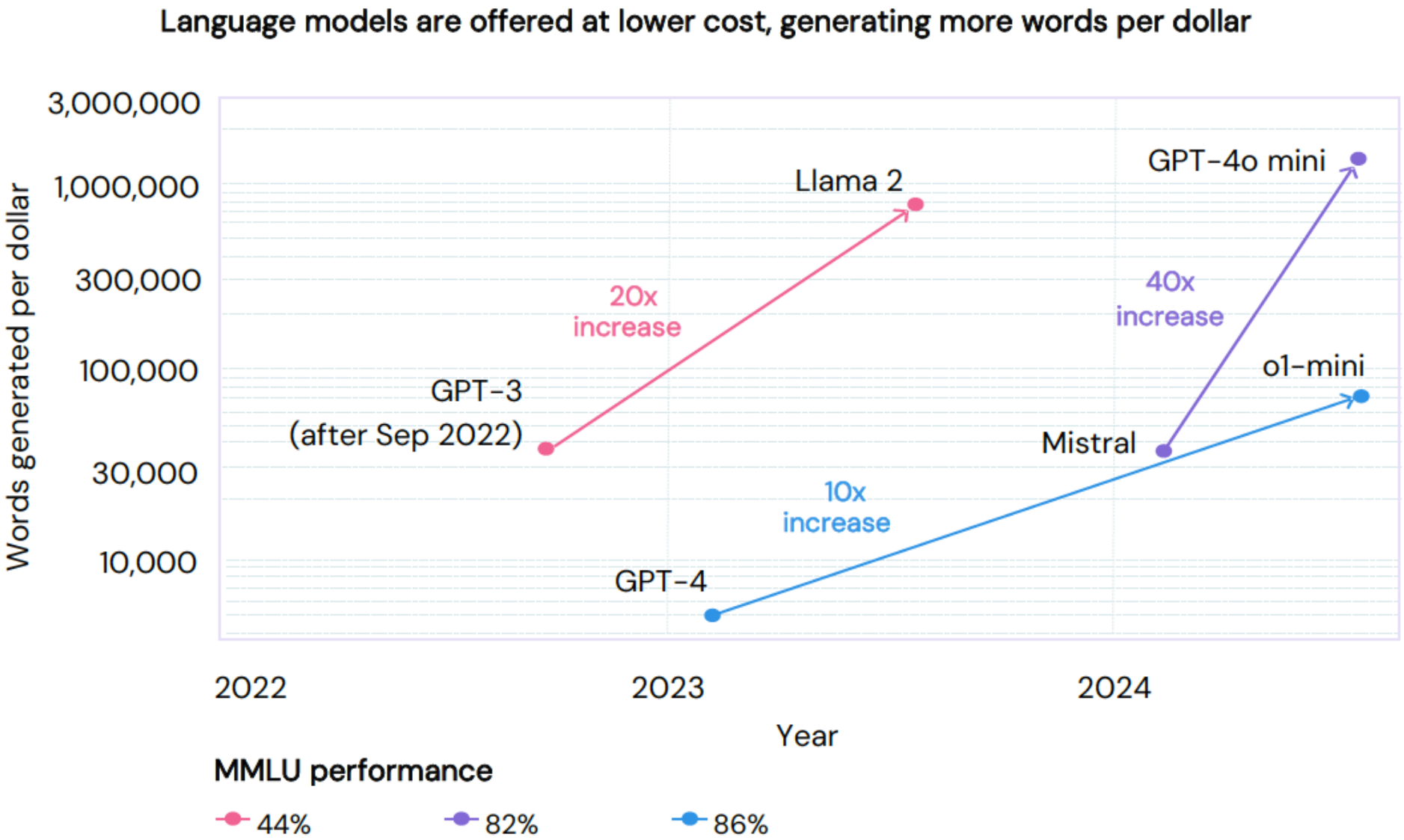
- Data availability, availability of GPU chips
- Energy, ‘capital’.

But there’s also scaling in a new type of agentic LLM system design, that supports action sequencing and reasoning.

Training compute times are going up. . .



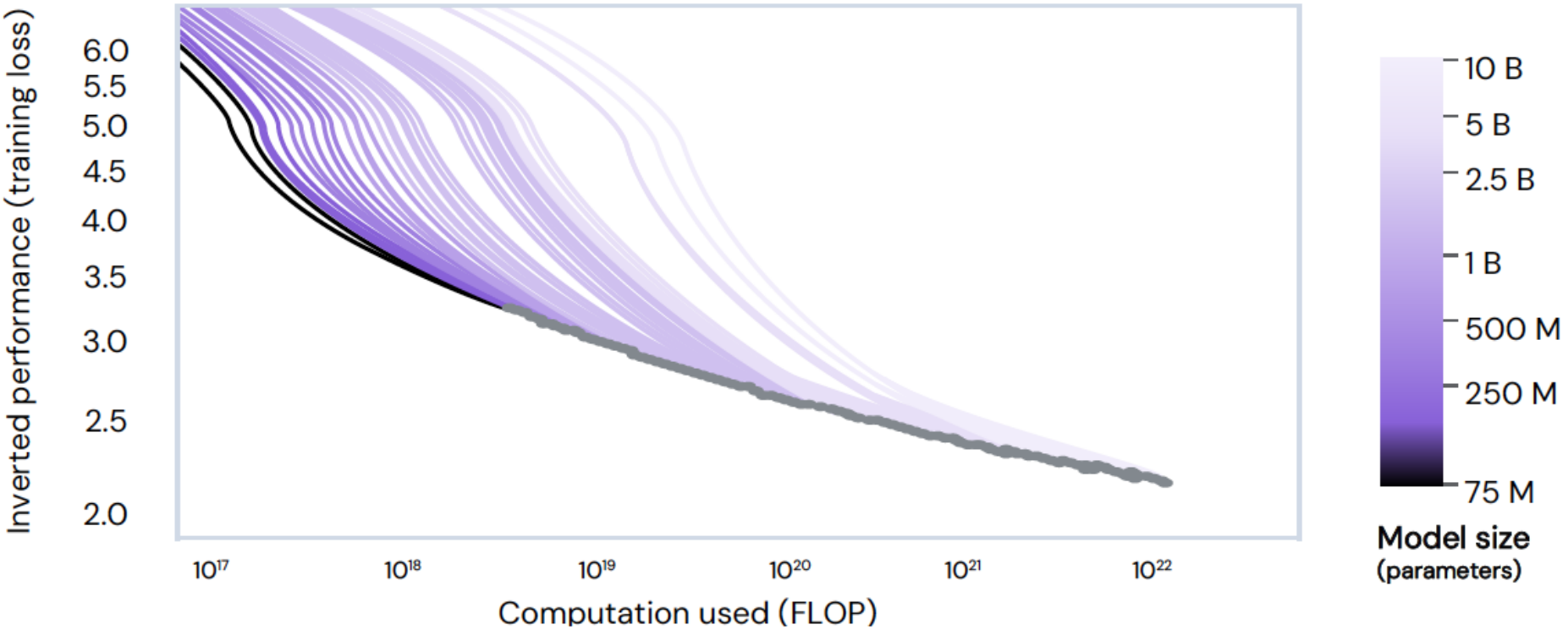
LLMs are becoming more efficient. . .



LLMs are becoming better ‘word predictors’…

Performance in ‘predicting the next word’ improves predictably with model size and training compute time.

- LLMs are (mainly) trained for this (kind of weird) ‘task’.



…but there are clearly diminishing returns…

What does the report say about future AI capabilities?

The report surfaces *disagreement*:

- 'Experts disagree on what to expect even in the coming months and years.'
- 'Experts variously support the possibility of general-purpose AI capabilities advancing slowly, rapidly, or extremely rapidly.'

There's some agreement about the 'remaining limitations of today's systems':

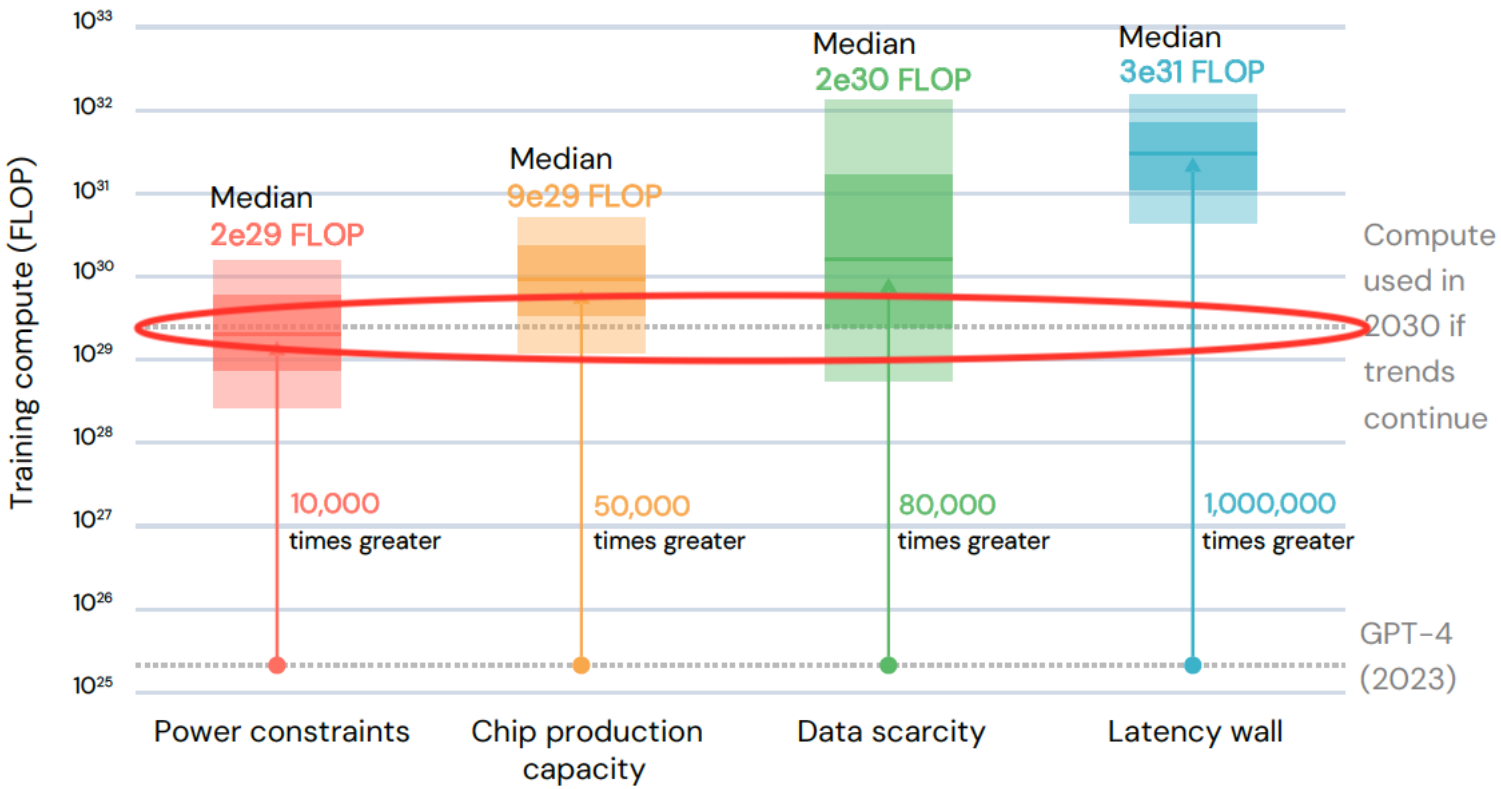
- 'Unreliability at acting in the physical world'
- 'Unreliability at executing extended tasks on computers'.

The main disagreement is in how far 'further scaling' will solve these problems.

What might limit increases in compute power?

This graph shows the compute we'll have in 2030, if current trends continue.

- The bars show estimated limits from different sources.
- This suggests we can continue at the current rate.



LLM agents

AI agents are LLMs that can 'autonomously make plans, perform complex tasks, and interact with their environment by controlling software and computers, with little human oversight'.

- Regular LLMs learn to generate words. . . agents also learn to generate 'actions' alongside words.
  - 'Actions' are performed with tools made available to the system.
  - E.g. a web browser, a computer command line, a robot.
- Example tasks include online shopping, assistance with scientific research, following instructions to navigate simulated environments, controlling physical robots.
- On these tasks, current AI agents mostly succeed in cases of low to medium complexity, but 'fail when the task requires many steps or becomes more complex'.
- Experts were divided about how scaling may help these systems.

LLM agents for reasoning: OpenAI's o1 model

o1 uses agent-like processing to perform *complex reasoning*.

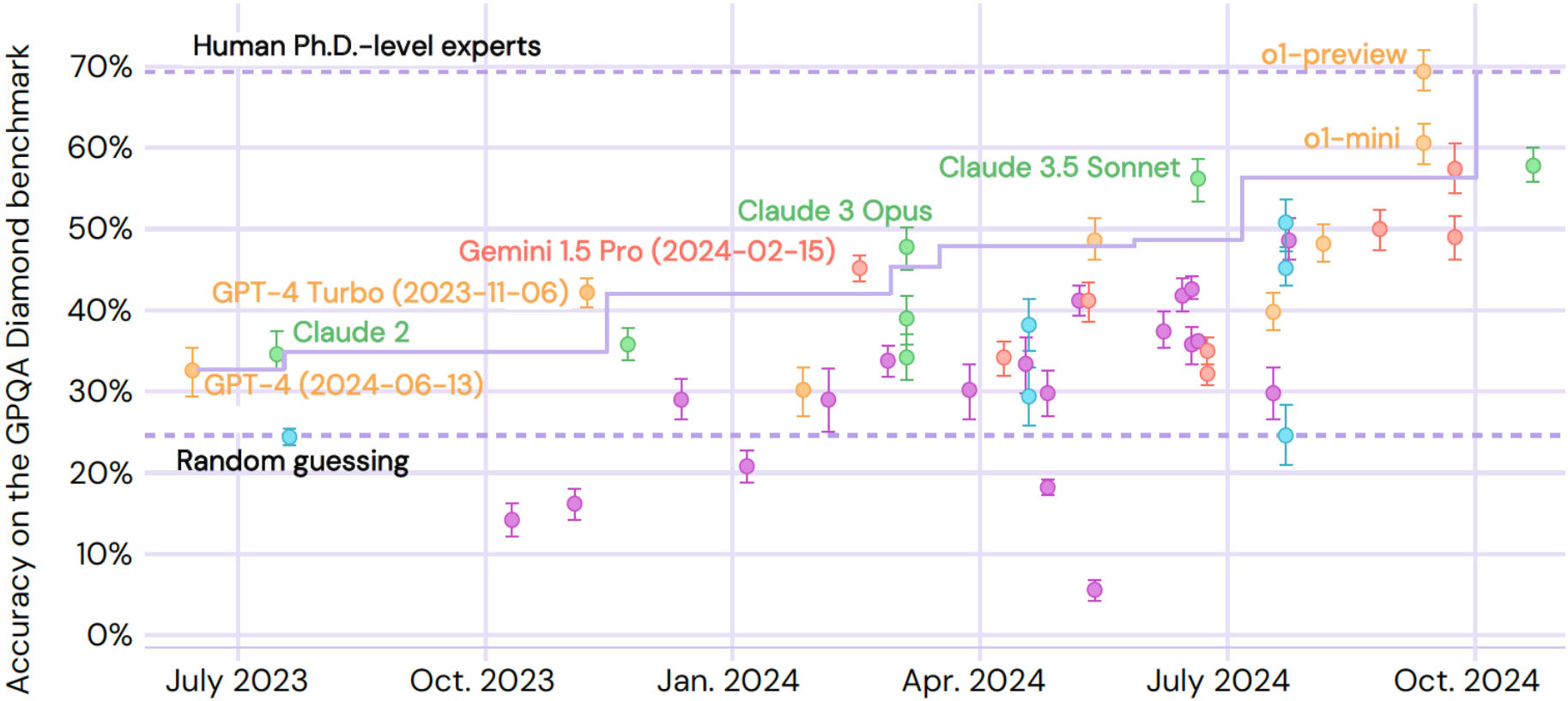
- It 'thinks' before it answers, producing a long internal chain of thought before responding to the user.

A LLM often works best if you break a task up into steps, and give it one step at a time.

- o1 automates the process of *breaking a task up into steps*: it does this part itself.
- Then it executes the steps it designed for itself.

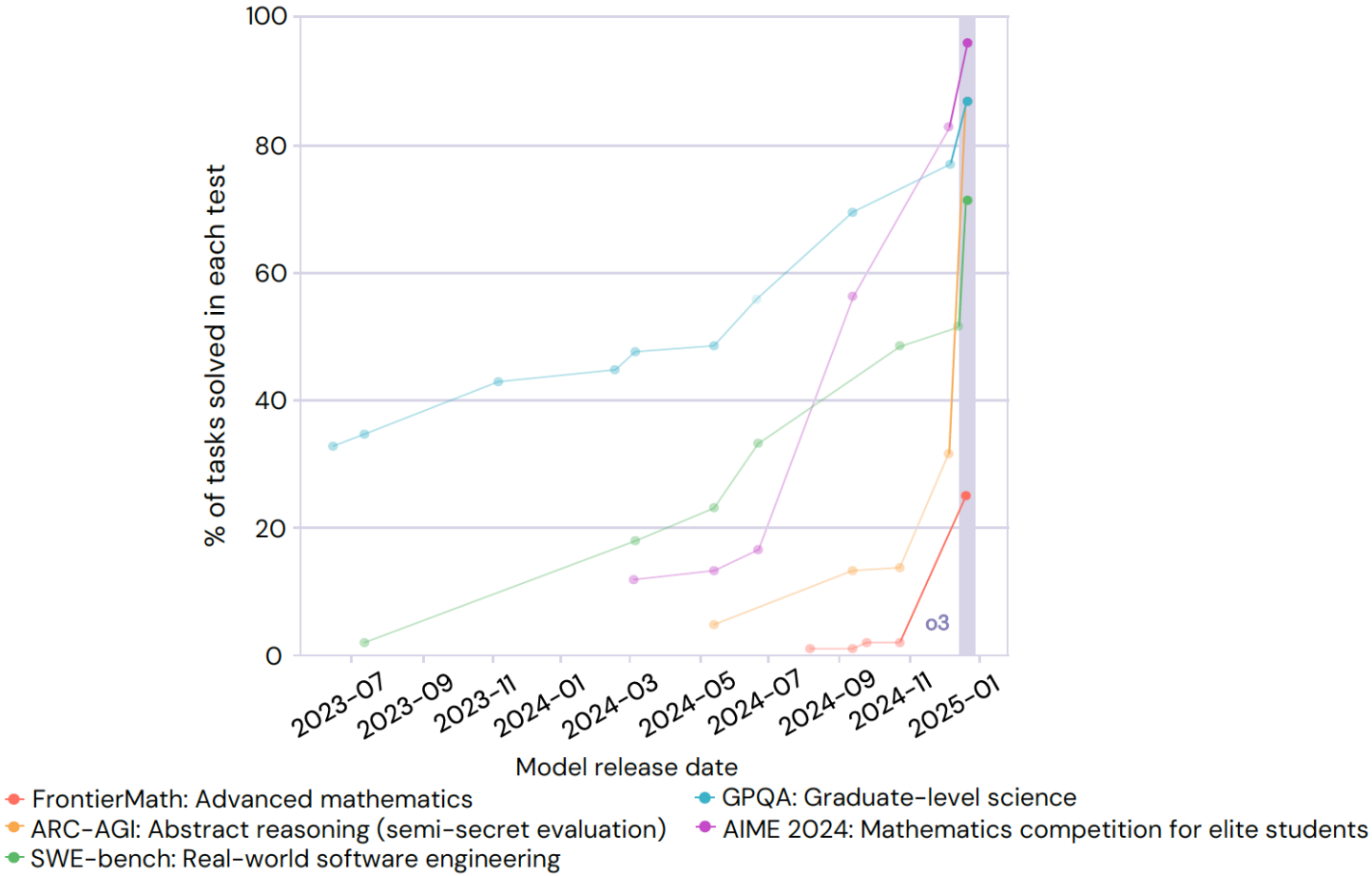
Is o1 at PhD-level task performance?

In September 2024, o1 qualified for the US Maths Olympiad, and ‘reached expert PhD-level performance on postgraduate-level physics, chemistry, and biology questions curated for high difficulty’.

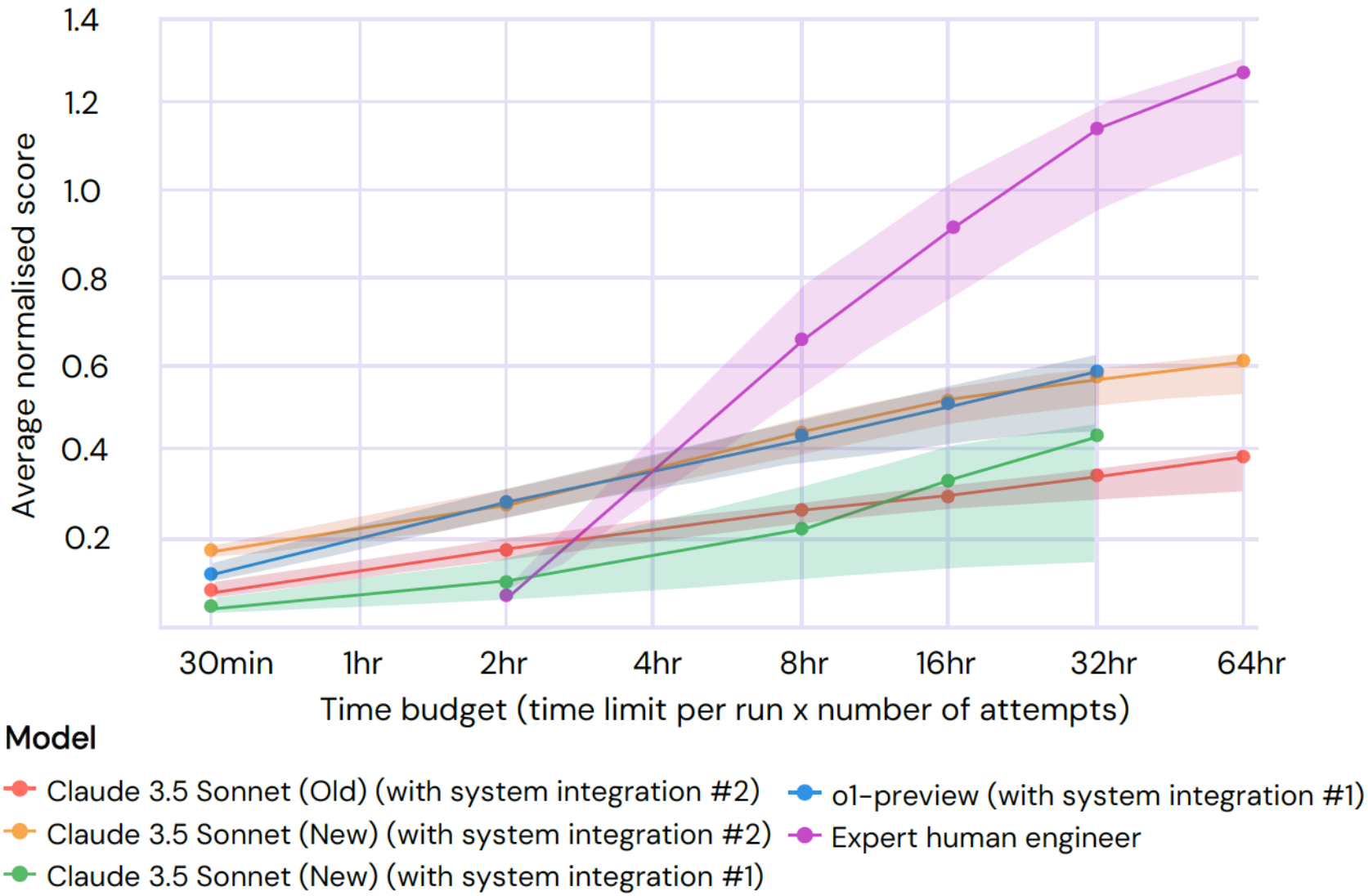


Stop press: o3

A last minute ‘Chair’s update’ was added to the report, presenting OpenAI’s o3 model. (Scaling seems to be effective here. . .)



Humans still better than AI, for tasks needing time. . .



2. What are the risks of current & future AI systems?

The report identifies three main categories of risk, for now and the future:

- Risks from malicious use. (AI used with harmful intent)
- Risks from malfunctions. (Unintentional harms arising from use)
- Systemic risks, arising not from 'use', but from side-effects.

2.1. Risks from malicious use of AI

2.1.1. Harm to individuals through fake content.

- Deepfake pornography, AI-generated child sexual abuse material
- Financial fraud through voice impersonation
- Blackmail for extortion; sabotage of reputation
- Bullying / psychological abuse.

Reports of these kinds of content are common, but reliable statistics on their frequency of these incidents are lacking.

2.1. Risks from malicious use of AI

2.1.2. Manipulation of public opinion.

- This is a harm *for society*, rather than for individuals.

AI ‘makes it easier to generate persuasive content at scale’.

- This can help subvert political (democratic) processes.
  - It can involve false content, but it doesn’t have to.
- However, evidence on how prevalent and how effective such efforts are remains limited.

2.1. Risks from malicious use of AI

2.1.3. Uses of AI in cyberattacks.

- AI can make it easier or faster for malicious actors of varying skill levels to conduct cyberattacks.
- New agentic AI seems to pose a few more risks here.

2.1.4. Uses of AI in chemical and biological attacks.

- Similar considerations here.

2.2. Risks from malfunctions of AI

2.2.1. Risks from AI failing to perform properly.

- For example: if users consult a general-purpose AI system for medical or legal advice, the system might generate an answer that is partly or completely wrong.
- Users are often not aware of the limitations of current AI products.
  - That's partly because of limited 'AI literacy' in current users...
  - Exacerbated by misleading advertising, and marketing pushes.
- Another problem is 'automation bias':
  - If a system normally works alright, it's hard for humans to monitor thoroughly.

2.2. Risks from malfunctions of AI

2.2.2. Risks from biased operation of AI systems.

- AI systems ‘frequently display biases with respect to race, gender, culture, age, disability, political opinion, or other aspects of human identity’.
  - Note, biased operation is classed as ‘malfunction’, rather than ‘malicious use’. I think that’s mostly right.
- Biased operation of AIs can lead to discriminatory outcomes: including ‘unequal resource allocation, reinforcement of stereotypes, and systematic neglect of underrepresented groups or viewpoints’.
  - Some harms arise because AIs make decisions about people. . .
  - Others arise just through dissemination of biased AI content.

2.2. Risks from malfunctions of AI

2.2.2. Risks from 'loss of control' (by humans, to AIs).

- This is a hypothesised *future* risk. A scenario in which 'one or more AI systems come to operate outside of anyone's control, with no clear path to regaining control'.
- There is broad consensus that current AI doesn't pose this risk.
- But experts are divided about how likely it will be 'within the next several years'.
  - Some consider it implausible, some consider it likely to occur...
  - Some see it as a modest-likelihood risk that warrants attention due to its high potential severity.
- Ongoing empirical and mathematical research is gradually advancing these debates.

2.3. ‘Systemic’ risks of AI

2.3.1. Labour market risks

- ‘Many people could lose their current jobs’ to AI.
- But ‘many economists expect that potential job losses could be offset, partly or potentially even completely, by the creation of new jobs and by increased demand in non-automated sectors’.

Stop press: new analyses of the labour market suggest AI is being taken up faster than previously thought, to do more tasks. . .

2.3. ‘Systemic’ risks of AI

2.3.2. Global AI R&D divide

- AI R&D is currently concentrated in ‘a few Western countries and China’.
- This could ‘increase the world’s dependence’ on these countries.
- ‘Some experts also expect it to contribute to global inequality.’
- ‘Access to compute’ is a particular problem for smaller countries.

2.3. ‘Systemic’ risks of AI

2.3.3. Market concentration and single points of failure

- A small number of companies currently dominate the AI market.
  - So there are only a few big AI systems.
- A failure in one of these ‘could cause simultaneous failures and disruptions on a broad scale’.

To my mind, ‘market concentration’ also creates economic inequalities, within countries as well as between. . .

2.3. ‘Systemic’ risks of AI

2.3.4. Environmental risks

- The amount of ‘energy, water, and raw material’ consumed by the AI industry is growing fast.
- There’s no sign of this growth slowing, despite various ‘new efficiencies’.
  - Remember this report predated DeepSeek. . .
  - But also remember the Jevons paradox!

2.3. ‘Systemic’ risks of AI

2.3.5. Privacy risks

- Private information from the training data can leak to a user. . .
- If users share private information in prompts, this can also leak.
  - E.g. into other training sets.
- Bad actors can also use AI to *infer* private info from datasets.
  - And Gen AI can do this kind of inference too.

‘So far, researchers have not found evidence of widespread privacy violations’ caused by AI.

- But those new labour market analyses may prompt a revised assessment. . .

2.3. ‘Systemic’ risks of AI

2.3.6. Copyright infringement risks

- AI training sets often contain lots of copyright material.
- Given legal copyright uncertainties, AI companies are ‘sharing less information’. . . which is a problem.

So what do you recommend??!

2.3. ‘Systemic’ risks of AI

2.3.7. Open-weights AI models: a separate dimension of systemic risk

- Open-weight models can pose risks, e.g. ‘by facilitating malicious or misguided use that is hard for the developer to monitor or mitigate’.
- Once weights are available for download, ‘there is no way to implement a rollback’. . .

The concept of marginal risk is useful here:

- Will releasing an open-weight model increase or decrease a given risk, relative to the risks of a closed model?

3. What techniques are there for managing AI risks?

The report surveys two things:

- Some factors which make AI risk management *particularly hard*
- Some techniques and frameworks that can (nevertheless) be used.

Within mitigation frameworks, regulation is (apparently) in scope.

- But it's not a big focus—there could be more in a report like this!
- The report explicitly sidesteps discussion of whether 'regulation will impede the speed of AI advances'.

3.1. What makes AI risk management specially hard?

3.1.1. A few technical features of AI:

- AI systems have an *unusually broad range of possible uses*.
  - LLMs can be used everywhere. . . so it's hard to anticipate risks.
- Developers 'still understand little about how their systems work'.
  - This makes it hard to *predict* problems, & also to *resolve* them.
  - AI system interpretability is a big (ongoing) open research field.
    - Some new progress in 'mechanistic interpretability'. . .
- 'Agentic' AIs 'present new challenges for risk management'.
  - These are agents which are 'out there in the world doing stuff'. We're *just starting* to think how to manage the new risks here.
  - Another issue: these systems are trained with reinforcement learning, which means they find *whatever means they can* to achieve their goal. Humans don't oversee the AI's *methods*. . .

3.1. What makes AI risk management specially hard?

3.1.2. A few ‘political and economic’ factors.

3.1.2.1. The rapid pace of some AI advances creates an ‘evidence dilemma’ for decision-makers.

- Progress can happen ‘in leaps’: so risks can also emerge quickly.
  - Example: the sudden new risks of AI cheating in schools.
- All risks are better managed pre-emptively.
  - The report says, especially for sudden risks? I’m not sure. . .
- In either case, it’s *hard to justify pre-emptive mitigation steps before there’s any evidence of a risk*.

To address the problem, companies and governments need very efficient ‘early warning systems’.

3.2. Techniques/frameworks for AI risk management

There are two things here:

- Methods for *identifying / monitoring* risks
- Methods for *mitigating* risks.

3.2.1. Methods for identifying / monitoring risks

Methods for *proactively identifying* risks involve running 'spot checks':

- Test a single AI system, in a set of specific tasks / contexts.
  - 'Red-teaming methods' are a common approach.
- But spot-checks can *miss a lot*. (Because AI systems are *general*.)
- 'No current method can reliably prevent unsafe outputs.'

Evaluators need a few things:

- 'Substantial technical expertise'
- 'Substantial resources'
- 'Sufficient access to relevant information'.

Okay, so ask for these things! The report should ask for the necessary structures, laws, resources...

3.2.1. Methods for identifying / monitoring risks

There are also ways of *monitoring* the inputs & outputs of AI systems, while they are in use, to look for harful content.

- But ‘moderately skilled users can often circumvent’ these methods.
- The methods also ‘introduce costs and delays’—especially if human oversight is involved.

3.2.1. Methods for identifying / monitoring risks

If we can identify AI-generated content, then we know *where to look* (for some risks).

The report discusses a few methods for detecting AI-generated content.

- It focusses on watermarking. . .
  - Again, 'moderately skilled users can circumvent' this method.
- The report could talk more about other detection methods!

The report suggests it might be more effective to use provenance tools to positively identify human-generated content.

- Our GPAI paper gets a nod here. . . that's our one mention.

3.2.1. Methods for identifying / monitoring risks

For privacy risks, we do know of some methods, that are somewhat effective, right through the AI development lifecycle.

- Removing sensitive information from training data
- ‘Differential privacy’ methods, that control how much information is learned from data
- ‘Privacy-enhancing’ techniques that make it hard to recover the system’s training data.

The report's conclusion

'The future of general-purpose AI is uncertain, with a wide range of trajectories appearing possible even in the near future, including both very positive and very negative outcomes'.

How AI is further developed, *who* develops it, and *which tasks* it is deployed for, all 'depend on the choices that societies make today, and in the future'.

My conclusion

The report is really good on the science & engineering, and on the political/economic problems surrounding AI safety discussions.

But it also misses opportunities to give governments *concrete advice* on regulatory options. It could have recommended:

- More on *mechanisms* for advancing AI safety.
  - E.g. AI Safety Institutes, new structures, new laws.
- Specific actions for Summit governments:
  - *Enforce laws* that provide needed transparency from companies
  - *Enact new laws* where they are needed
  - *Upskill staff* in enforcement agencies
  - *Improve funding* of existing enforcement agencies.

Perhaps it's not surprising the Paris communiqué didn't refer to it :-