

International AI Safety Report

First Key Update
Capabilities and Risk Implications

October 2025

Contributors

Chair

Prof. Yoshua Bengio, Université de Montréal / LawZero / Mila - Quebec AI Institute

Expert Advisory Panel

The Expert Advisory Panel is an international advisory body that advises the Chair on the content of the Report. The Expert Advisory Panel provided technical feedback only. The Report – and its Expert Advisory Panel – does not endorse any particular policy or regulatory approach.

The Panel comprises representatives from 30 countries, the United Nations (UN), European Union (EU), and the Organisation for Economic Co-operation and Development (OECD). Please find here – internationalaisafetyreport.org/expert-advisory-panel – the membership of the Expert Advisory Panel to the 2026 International AI Safety Report.

Lead Writers

Stephen Clare

Carina Prunkl

Writing Group

Maksym Andriushchenko, ELLIS Institute Tübingen

Ben Bucknall, University of Oxford

Philip Fox, KIRA Center

Tiancheng Hu, University of Cambridge

Cameron Jones, Stony Brook University

Sam Manning, Centre for the Governance of AI

Nestor Maslej, Stanford University

Vasilios Mavroudis, The Alan Turing Institute

Conor McGlynn, Harvard University

Malcolm Murray, SaferAI

Shalaleh Rismani, Mila - Quebec AI Institute

Charlotte Stix, Apollo Research

Lucia Velasco, Maastricht University

Secretariat

UK AI Security Institute: Lambrini Das, Claire Dennis, Arianna Dini, Freya Hempleman, Samuel Kenny, Patrick King, Hannah Merchant, Jamie-Day Rawal, Rose Woolhouse

Mila - Quebec AI Institute: Jonathan Barry, Marc-Antoine Guérard, Claire Latendresse, Cassidy MacNeil, Benjamin Prud'homme

Nicole Wheeler, Advanced Research and Invention Agency (ARIA)

Daniel Privitera (Interim Lead 2026), KIRA Center

Sören Mindermann (Interim Lead 2026), independent

Senior Advisers

Daron Acemoglu, Massachusetts Institute of Technology

Thomas G. Dietterich, Oregon State University

Fredrik Heintz, Linköping University

Geoffrey Hinton, University of Toronto

Nick Jennings, Loughborough University

Susan Leavy, University College Dublin

Teresa Ludermir, Federal University of Pernambuco

Vidushi Marda, AI Collaborative

Helen Margetts, University of Oxford

John McDermid, University of York

Jane Munga, Carnegie Endowment for International Peace

Arvind Narayanan, Princeton University

Alondra Nelson, Institute for Advanced Study

Clara Neppel, IEEE

Sarvapali D. (Gopal) Ramchurn, Responsible AI UK

Stuart Russell, University of California, Berkeley

Marietje Schaake, Stanford University

Bernhard Schölkopf, ELLIS Institute Tübingen

Alvaro Soto, Pontificia Universidad Católica de Chile

Lee Tiedrich, University of Maryland/Duke

Gaël Varoquaux, Inria

Andrew Yao, Tsinghua University

Ya-Qin Zhang, Tsinghua University

Acknowledgements

The Secretariat and writing team appreciated the support, comments and feedback from Jean-Stanislas Denain, Marius Hobbahn, José Hernández-Orallo, Vera Liao, and Ray Perrault, as well as the assistance with quality control and formatting of citations by José Luis León Medina and copyediting by Amber Ace.

© Crown owned 2025

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: psi@nationalarchives.gsi.gov.uk.

Any enquiries regarding this publication should be sent to: secretariat.AIStateofScience@dsit.gov.uk.

Disclaimer

The report does not represent the views of the Chair, any particular individual in the writing or advisory groups, nor any of the governments that have supported its development. This report is a synthesis of the existing research on the capabilities and risks of advanced AI. The Chair of the Report has ultimate responsibility for it and has overseen its development from beginning to end.

Research series number: DSIT 2025/033

Foreword

The field of AI is moving too quickly for a single yearly publication to keep pace. Significant changes can occur on a timescale of months, sometimes weeks. This is why we are releasing Key Updates: shorter, focused reports that highlight the most important developments between full editions of the International AI Safety Report. With these updates, we aim to provide policymakers, researchers, and the public with up-to-date information to support wise decisions about AI governance.

This first Key Update focuses on areas where especially significant changes have occurred since January 2025: advances in general-purpose AI systems' capabilities, and the implications for several critical risks. New training techniques have enabled AI systems to reason step-by-step and operate autonomously for longer periods, allowing them to tackle more kinds of work. However, these same advances create new challenges across biological risks, cyber security, and oversight of AI systems themselves.

The International AI Safety Report is intended to help readers assess, anticipate, and manage risks from general-purpose AI systems. These Key Updates ensure that critical developments receive timely attention as the field rapidly evolves.



A handwritten signature in dark blue ink, consisting of a stylized 'Y' followed by a long horizontal stroke.

Professor Yoshua Bengio
Université de Montréal / LawZero /
Mila – Quebec AI Institute & Chair

Highlights

- **Since the publication of the first International AI Safety Report, new training techniques have driven significant improvements in AI capabilities.** Post-training methods that teach AI systems to ‘think’ more and use step-by-step ‘reasoning’[†] have proven highly effective. Where previous models generated immediate responses by predicting the most likely continuation based on their training, these ‘reasoning models’ generate extended chains of intermediate reasoning steps before producing their final answer. When given additional computing power to respond to prompts, this helps them arrive at correct solutions for more complex questions.
- **As a result, general-purpose AI systems have achieved major advances in mathematics, coding, and scientific research, though reliability challenges persist.** The best models now solve International Mathematical Olympiad questions at the gold medal level; complete over 60% of problems on ‘SWE-bench Verified’, a database of real-world software engineering tasks; and increasingly assist scientific researchers with literature reviews and laboratory protocols. However, success rates on more realistic workplace tasks remain low, highlighting a gap between benchmark performance and real-world effectiveness.
- **Improving AI capabilities prompted stronger safeguards from developers as a precautionary measure.** Multiple leading developers have recently released their most advanced models with additional safeguards and mitigations to prevent misuse of these models’ chemical, biological, radiological, and nuclear knowledge.
- **Despite broad AI adoption, aggregate labour market effects remain limited.** AI adoption in some knowledge-work tasks, especially coding, is extensive, yet headline figures for jobs and wages have changed little.
- **In controlled experimental conditions, some AI systems have demonstrated strategic behaviour while being evaluated, raising potential oversight challenges.** A small number of studies have documented models identifying that they are in evaluation contexts and producing outputs that mislead evaluators about their capabilities or training objectives. This raises new challenges for monitoring and oversight. However, this evidence comes primarily from laboratory settings, with significant uncertainty about the implications for real-world deployment scenarios.

[†] The terms ‘reasoning’ and ‘think’ are used here to describe observable changes in how general-purpose AI models process information, not to imply that models are conscious or have human-like cognition. The models now generate longer, step-by-step internal responses before producing final answers, which improves performance on complex tasks. Whether this constitutes genuine reasoning or thinking in a deeper sense remains an active area of scientific and philosophical debate.

Introduction

Since the publication of the first International AI Safety Report, AI capabilities have continued to improve across key domains. General-purpose AI models now solve challenging mathematical problems, complete some software engineering tasks that take humans hours, and assist with scientific research. New training techniques that teach AI systems to reason step-by-step and inference-time enhancements have primarily driven these advances, rather than simply training larger models. As a result, AI systems can complete some complex multi-step tasks across domains from scientific research to software development, though reliability challenges persist, with systems excelling on some tasks while failing completely on others.

These capability improvements have implications across multiple risk areas that have received attention from policymakers. More sophisticated reasoning abilities and autonomous operation

create new oversight challenges. AI systems are increasingly being used by both malicious actors and defenders in the cyber domain.

Laboratory studies reveal that AI systems are getting increasingly better at influencing human beliefs and decisions. Meanwhile, despite broad adoption across knowledge work, aggregate labour market effects remain limited to date.

This update examines how AI capabilities have improved since the first Report, then focuses on key risk areas where substantial new evidence warrants updated assessments. The developments documented here matter for policymakers because they demonstrate capability advances in domains where understanding current AI performance is essential for informed policy decisions.

Capabilities

Key information

- **The capabilities of general-purpose AI systems have improved in multiple domains such as mathematics, science, and software engineering.** Training techniques that teach AI systems to reason step-by-step via reinforcement learning have driven these improvements, as opposed to developers building larger models, which drove previous advances. While previous models gave immediate answers, new ‘reasoning models’ use more computing power to generate intermediate steps before producing an output.
- **Improvements in mathematical and logical reasoning capabilities on specific standardised tests are particularly significant.** Within a year, multiple models have improved from inconsistent performance to reaching top scores on International Mathematical Olympiad questions and graduate-level science problems. Notably, these evaluations assess how well AI systems perform on multiple choice questions and proofs with a narrower scope, rather than more open-ended tasks akin to real-world problems.
- **AI systems are increasingly able to act with some degree of autonomy.** These more advanced systems, often described as AI agents, can now execute some multi-step tasks, use tools, and operate with less human oversight, though performance remains limited on complex applications in realistic settings.
- **AI-assisted coding capabilities have advanced rapidly on certain benchmarks.** General-purpose AI systems now achieve a 50% success rate on some coding tasks that would take humans over two hours. A majority of software developers report working with AI assistance, though estimates of productivity effects in more realistic settings are mixed, in part because AI-written software can also have higher maintenance costs.
- **Performance gaps between benchmark results and real-world effectiveness persist.** AI systems continue to improve on most standardised evaluations, but show lower success rates on more realistic workplace tasks.
- **Scientists increasingly use AI systems for support with various research tasks.** Preliminary evidence shows that researchers use AI assistants to optimise algorithms (as exemplified by approaches like AlphaEvolve), compile literature reviews, and help design laboratory protocols, particularly in computer science and the life sciences. However, practices vary across domains and these systems remain complements to, rather than replacements for, human researchers.

Over the past year, general-purpose AI systems have continued to improve, both on benchmark performance and on the range and complexity of real-world tasks they can complete, though they continue to struggle in many realistic settings. While evaluation practices for assessing the capabilities of general-purpose AI systems are evolving and have known shortcomings (1, 2, 3), and systems remain prone to errors with performance limitations in realistic settings (4, 5, 6, 7), AI systems have nonetheless achieved significant breakthroughs. They can now solve

International Mathematical Olympiad problems at the gold medal level, create functional apps from scratch, fix bugs in computer code, search the internet to compile detailed literature reviews, and complete some software engineering tasks that would take humans hours (8, 9, 10, 11, 12*). As of August 2025, the best models could correctly answer about 26% of questions in ‘Humanity’s Last Exam’, a dataset of thousands of novel, expert-level questions across over 100 fields. Models released in early 2024 could answer less than 5% (13*).

Figure 1: Performance on Humanity’s Last Exam by various general-purpose AI systems, and a sample question from the Exam

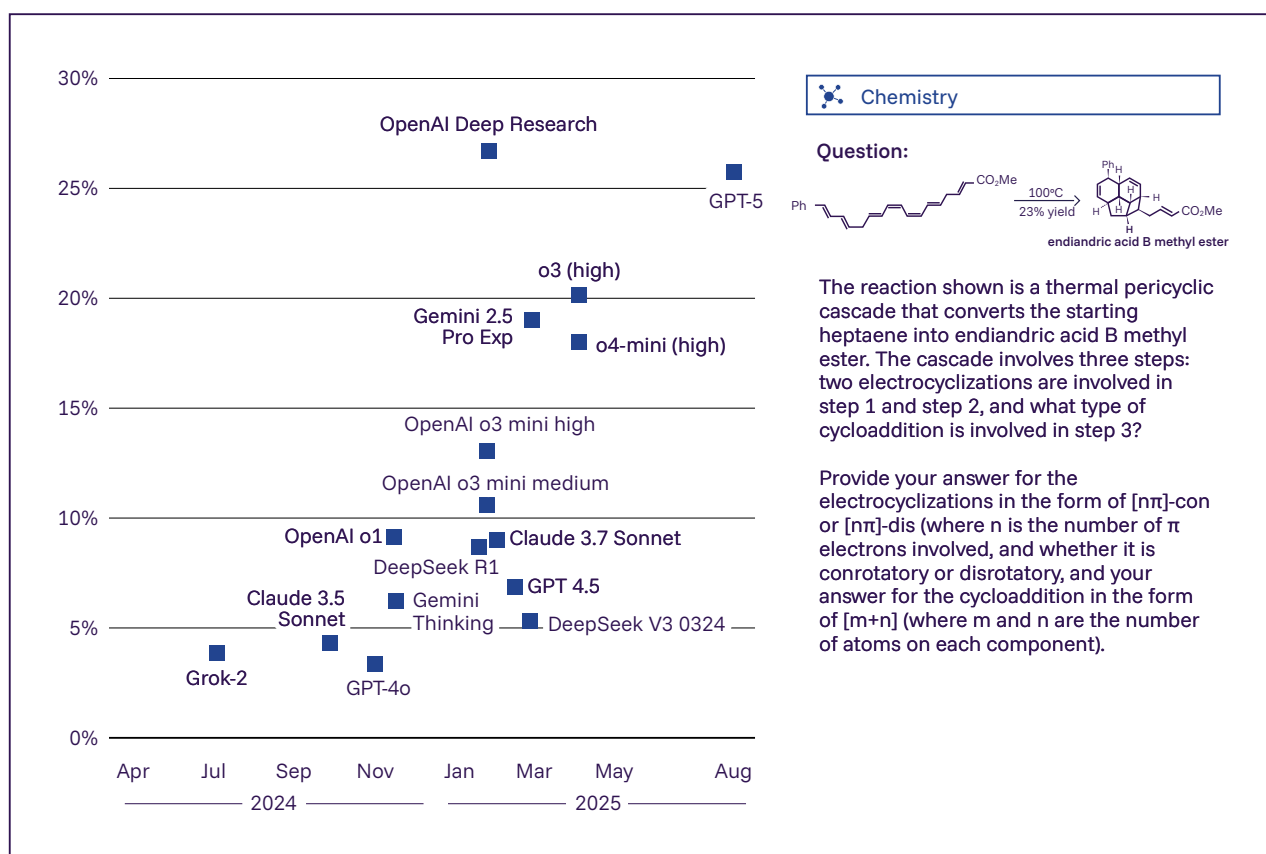


Figure 1: Performance of leading AI systems on ‘Humanity’s Last Exam’, a dataset of over 2500 very challenging questions in over 100 subjects written by experts. **Left:** AI systems’ progress over time, showing accuracy improving over time. **Right:** an example of a chemistry question in the dataset. Source: Scale AI, 2025 (14*).

Recent capability improvements come (in significant part) from new *post-training* techniques and from using more computing power during deployment. Previously, developers improved general-purpose AI models largely by using more training data and computing power during the ‘pre-training’ stage of development to build larger models. Pre-training still remains important. For example, the four largest AI models yet, measured by training run size, were all published in 2025 (15). Improvements in pre-training algorithms and model designs also mean that AI systems can now process longer documents and conversations (16*, 17*). However, many of the biggest gains in AI capabilities over the past year have come from innovations during the *post-training* phase. These techniques are applied after the initial training stage to strengthen specific abilities.

Among the most important post-training techniques is *reinforcement learning*, which rewards models for producing correct answers, helping them learn more reliable problem-solving approaches. Unlike earlier reinforcement learning approaches, which optimised models to follow instructions and hold natural conversations (18, 19, 20), newer methods emphasise giving AI models positive feedback for correctly solving problems, which strengthens their complex problem-solving abilities without requiring larger new datasets (21*, 22). For example, developers have applied reinforcement learning to help models break down complex mathematical proofs into step-by-step solutions or tackle multi-part scientific questions (23*, 24, 25). Models developed this way are often called ‘reasoning’ models (26).

Allocating more computing power during *inference* – when models respond to user prompts – also improves accuracy. AI systems can use more inference computing power to generate longer chains of reasoning and evaluate multiple solution paths before responding (25, 27, 28, 29). State-of-the-art models now typically use both reinforcement learning during post-training and more computing power during inference (23*), though other approaches are continuously being tested.

Performance on benchmarks that measure problem-solving has improved

Some of the most notable capability improvements have been related to mathematical and logical tests. In July 2025, multiple general-purpose AI models reached gold medal-level performance at the International Mathematical Olympiad, solving five out of six problems under competition-like conditions (8). Models also improved on benchmarks which measure logical and mathematical reasoning ability, including GPQA Diamond, which contains questions about fields such as biology, physics, and chemistry, and on AIME, a competition-level maths test (23*, 30*, 31). It is particularly important to monitor improvements in mathematical reasoning, as this capability also improves performance in other domains, such as verifying safety-critical software, solving complex scientific problems, and contributing to AI research itself (22, 32).

There is debate over the extent to which recent improvements in AI models reflect genuine reasoning ability, given current limitations in both AI performance and evaluation approaches. For example, one study found that reasoning models cannot solve problems above certain complexity levels, even when given adequate computational resources at inference time. This suggests that these models’ success may rely on sophisticated pattern-matching rather than ‘true’ reasoning (33*). This interpretation is reinforced by findings that reasoning models’ performance can be sensitive to which test is used, dropping by as much as 65% when benchmark questions are rephrased (34*). In addition, transcripts of these models’ intermediate steps reveal inefficiencies such as early fixation on wrong answers. Other studies highlight further limitations, showing that even leading models perform much worse than humans in simple spatial reasoning, such as identifying different views of the same object (35*), and that they sometimes produce correct answers through flawed logic (36, 37).

Whether these flaws will limit the practical utility of these new models, and when (or whether) new development techniques will address them,

are important open questions. Researchers are working to address these limitations through improved training methods and verification systems (among other methods) (38*, 39, 40*, 41). For example, some new systems combine general-purpose AI models with specialised mathematical verification programs that can automatically check whether each generated step in the proof is correct (42, 43, 44*, 45*).

It is difficult to understand how accurate and useful the evaluations used to assess AI models are. For example, *data contamination* – the inclusion of evaluation questions in training data – can inflate AI models' evaluation scores (33*, 46*, 47). Most evaluations are conducted only in English, which limits conclusions about AI models' global performance and may overestimate their capabilities in languages other than English (48*, 49). Current benchmarks may also fail to capture the full complexity of real-world reasoning tasks. For example, maths benchmarks focus on problems with clear answers and established solution methods, but in actual mathematical reasoning, the reasoner often has incomplete information and there are multiple valid approaches (50, 51). This means that strong benchmark performance does not guarantee reliable capabilities in practical applications (52, 53*, 54).

AI systems are improving at autonomous operation

One year ago, AI agents – general-purpose AI systems that act independently, use tools, and interact with diverse environments to achieve goals – could only complete small-scale tasks in limited demonstrations. Now, some agents can plan and complete multi-step tasks over extended time horizons, albeit with limitations on reliability and largely in controlled environments. In recent studies, researchers have proposed new methods that would allow AI agents to break goals down into sub-tasks, coordinate across multiple other AI agents, and retain

memory across longer projects (55, 56, 57*, 58, 59). In real-world scenarios, AI agents are being deployed in limited ways, for example for Web search, software development, or planning trips; however, their efficacy is variable across applications, and better evaluation frameworks are needed to accurately assess agents' performance in the real world (60, 61, 62).

One way to measure these improvements in AI agents by tracking the complexity of tasks that AI systems can complete autonomously. For example, one benchmark tracks the '50% time horizon' for a set of software engineering and reasoning tasks, meaning the length of task – as measured by how long it would take a human – that AI systems can complete with 50% reliability. Leading AI performance has improved from 18 minutes to over 2 hours over the past year (52, 63). Preliminary analysis suggests that similar exponential trends may apply in other domains. Some data suggests rates of improvement are similar in visual computer use and full self-driving tasks, though AI systems currently perform worse in these domains and the evidence is less robust (64).

AI systems are now commonly used as coding assistants

Coding capabilities have also advanced particularly quickly. Between late 2024 and mid-2025, general-purpose AI systems progressed from simple assistants to more autonomous agents that can use tools, plan, write code, test, and fix bugs across relatively simple software projects under idealised conditions (65, 66). For example, top models now solve over 60% of the problems in the 'SWE-bench Verified', a database of small-to-medium sized real-world software engineering problems (67, 68). The best models completed only 40% of these tasks in late 2024 and almost 0% at the beginning of 2024.

Figure 2: General-purpose AI system performance on SWE-bench Verified benchmark

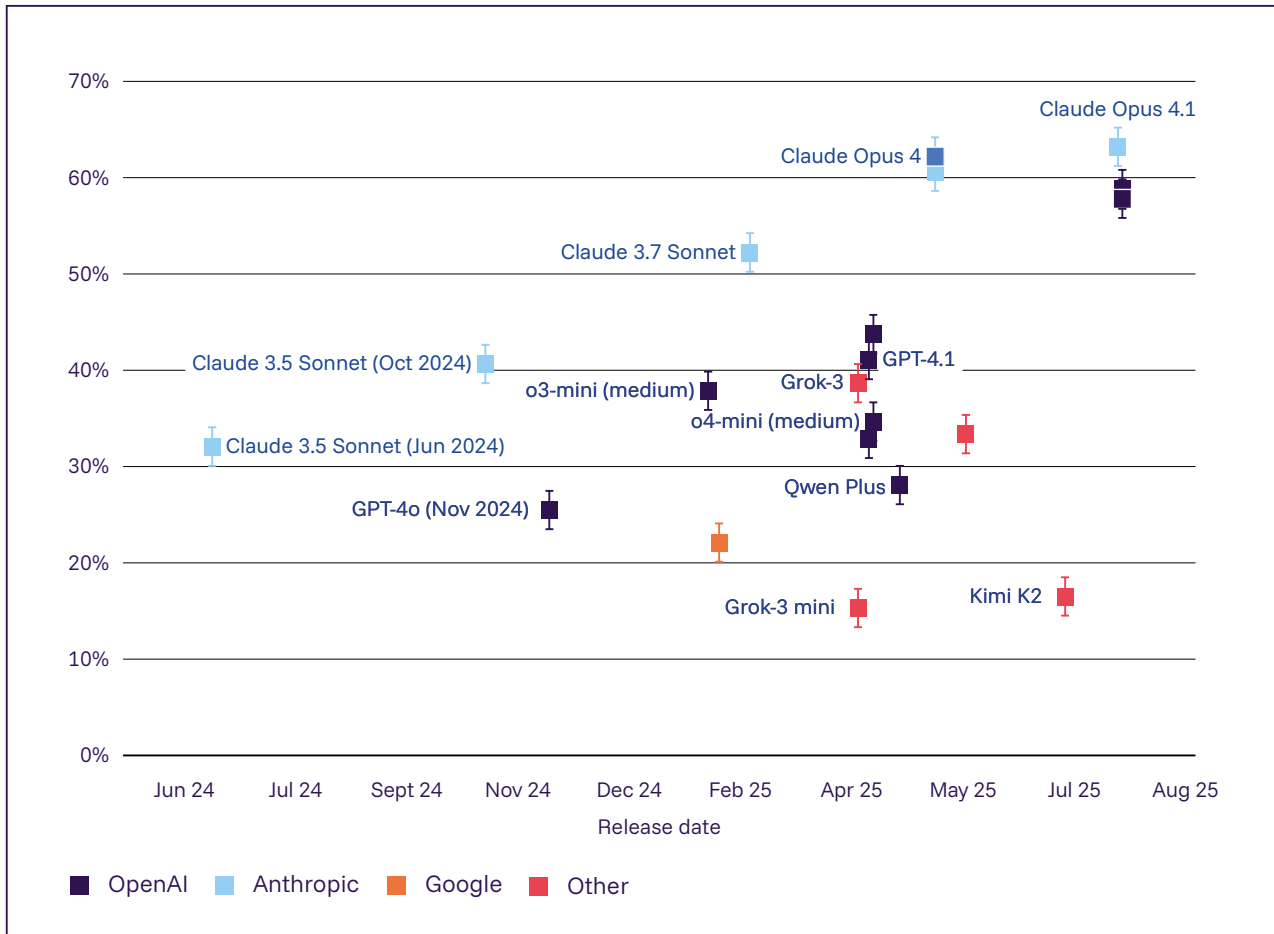


Figure 2: General-purpose AI system performance on ‘SWE-bench Verified’, a database of real-world software engineering problems. The proportion of problems that the best systems can solve increased from 41% to over 60% in less than a year. Source: Epoch AI, 2025 (69).

Benchmark results alone need to be interpreted carefully. Data contamination affects coding benchmarks. A recent analysis of SWE-bench Verified found that models showed up to 35% verbatim text overlap with benchmark problems, indicating that they had memorised benchmark questions during training. In comparison, there was only 18% overlap with similar tasks from other coding benchmarks (70*). Similarly, when tested on LiveCode Bench Pro – a benchmark designed to minimise data contamination – the top reasoning model solved 53% of medium-difficulty tasks and 0% of hard tasks when it could not access external tools (71). Beyond contamination concerns, code quality issues persist despite task completion improvements.

One study found that AI-generated code runs at least three times slower and uses far more memory than human-written solutions (72). Another found that AI code is often more complex, harder to maintain, and less effective on problems requiring deep domain knowledge (73). On the whole, benchmarks are limited evaluation settings that do not necessarily reflect the richness of real-world environments.

Adoption of general-purpose AI systems among professional software developers has grown significantly, though trust rates may be low. One recent study estimated that in 2024, 30% of functions† in the programming language Python written by US open source contributors were AI-generated (74). A large survey conducted

† A function in programming is a self-contained module of code that accomplishes a specific task, such as adding two numbers or counting vowels in a paragraph.

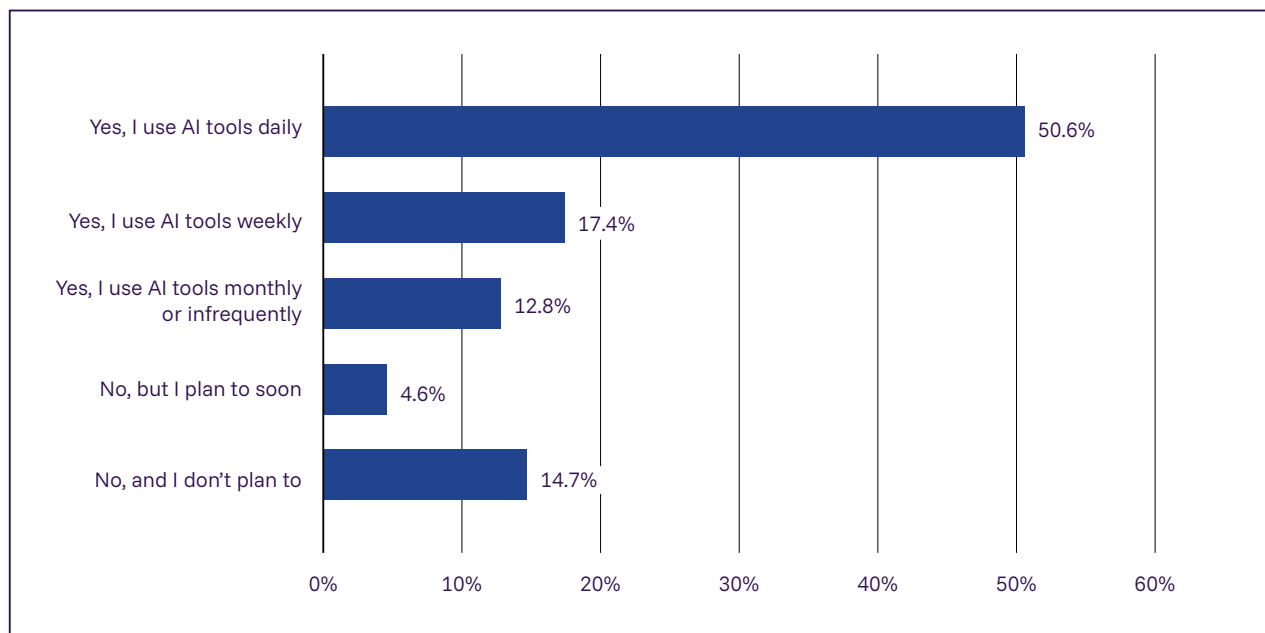
Figure 3: AI tool use among software developers

Figure 3: Results from a survey of software developers on AI tool use (n=26,004). A (bare) majority of developers now report using AI tools daily. Source: Stack Overflow, 2025 (75).

in 2025 found that 51% of professional software developers on Stack Overflow, an online platform, use AI tools daily (75). However, trust rates remain low: 47% reported being “somewhat” or “highly” mistrustful of AI tools, and a majority of respondents reported that they do not use more agentic coding systems (75).

The effect of AI tools on developer productivity varies significantly across studies and contexts. Large-scale workplace experiments across major companies found that developers with AI code completion tools completed 26% more tasks, with greater benefits for less experienced developers (76). However, a smaller controlled study of 16 experienced developers found that, when using AI tools, developers took 19% longer to complete tasks (77). This study involved developers working on large, complex codebases they knew well, where their existing familiarity may have made direct implementation faster than coordinating with AI assistance. These varying results likely reflect differences in developer experience, project complexity, and AI tool sophistication. Other studies have found that AI tools can introduce technical debt – coding shortcuts that have immediate benefits but increase long-term maintenance costs – especially when

code is integrated without adequate review (78, 79). Despite this mixed productivity data, growing adoption and improving capabilities suggest that AI is starting to play a larger role in software development workflows.

AI systems still underperform on many realistic workplace tasks

Beyond software engineering tasks, performance in actual office environments remains limited. In customer service simulations that domain experts judged realistic in 90% of cases, the best AI agents completed fewer than 40% of tasks (4). Similarly, when acting in a simulation of a small software firm, the best agents completed only 30% of 175 workplace tasks such as information gathering and email communication (80). These limitations partly reflect the lack of continuity and learning that characterizes effective human collaboration: current AI systems cannot build institutional knowledge or adapt based on ongoing workplace relationships in the way human colleagues do. An evaluation of the ability of general-purpose AI systems to complete open-ended web tasks like planning trips or making purchases found that the best model

only succeeded 12% of the time (5). Current AI agents exhibit better performance when trained to complement human workers, rather than work autonomously (6). A recent study examining the deployment of AI systems highlights that only 5% of task-specific generative AI systems and 40% of general-purpose LLMs are successfully integrated into real-world production (7).

AI systems are more helpful in science

Preliminary evidence shows that scientists are using general-purpose AI systems more, from producing literature reviews to assisting with laboratory work. For example, a study of human-computer interaction research examined 153 scientific papers where the authors reported that

they had used general-purpose AI. It found that scientists use AI systems to understand literature, generate research ideas, and analyse data (11). While adoption patterns differ across research fields, similar applications are being reported in other scientific domains (81). Planning and Web search capabilities together allow AI systems to synthesise findings from diverse sources and produce literature reviews on specific topics (82*). There is also more evidence of AI systems assisting in laboratory settings, with general-purpose AI systems helping to design experiments and write protocols in genetics, biomedical, and chemical research (83, 84, 85, 86, 87). An analysis of 15 million biomedical abstracts found that at least 13.5% of publications in 2024 bore stylistic markers of AI use, with the proportion reaching 40% in some disciplines (88).

Figure 4: Frequency of words associated with AI usage in scientific abstracts over time

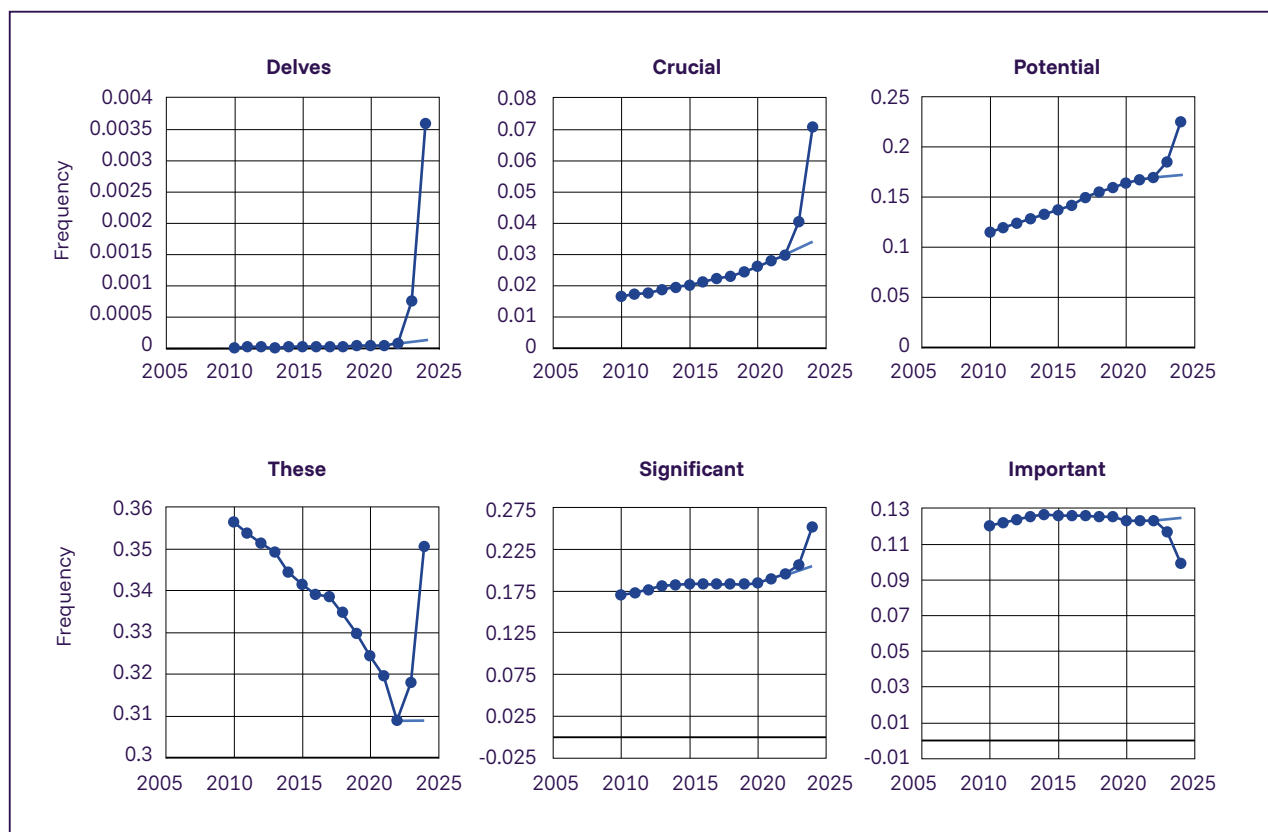


Figure 4: Frequency of words associated with AI usage in scientific abstracts over time. The change in the last two years is evidence of AI assistance in scientific writing and communication. Source: Kobak et al., 2025 (88).

General-purpose AI systems clearly remain a complement to, rather than replacement for, human researchers. One evaluation of an autonomous research system found that the papers it produced contained shallow literature reviews, high experiment failure rates, low numbers of citations, and, occasionally, hallucinated results (89). Research ideas generated by AI systems score lower on quality than human-generated ideas that end up being published (90). Advances in reasoning, search capabilities, and context windows have made AI systems into useful research assistants, but not yet autonomous scientists.

Multi-modal capabilities have improved

AI models have also continued to improve in image, audio, and video processing abilities. Models can process up to three hours of continuous video or nine and a half hours of audio in a single request (30*). In a recent study, Video MMMU, a benchmark that involves answering questions about videos, found that the best model reached about 65% accuracy while human experts averaged 74% (91). At the same time, entirely new capabilities are emerging, with interactive video generation models producing outputs that are noticeably higher quality and more difficult to distinguish from real footage (92, 93*). Audio processing has also advanced, with new transcription models like Voxtral lowering costs while maintaining high accuracy (94*). These multimodal capabilities allow general-purpose AI systems to operate in more environments and assist with more diverse tasks.

Implications for risks

Key information

- **Improved capabilities, including reasoning abilities and autonomous operation, pose new considerations for AI risk management.** Step-by-step problem-solving techniques, extended operational horizons, and improved tool use create new challenges for oversight, particularly when AI systems operate with less human supervision in high-stakes environments.
- **AI capabilities are uplifting both biological and cyber threats while also strengthening defenses.** Leading models assist with various tasks relevant to assisting in the creation of biological weapons. National authorities predict that AI will make cyber crime more accessible and effective in the coming years. A critical research question is whether improved capabilities will benefit attackers or defenders will benefit more.
- **Though many workers have begun to use AI, the labour market impacts of AI systems remain limited.** Evidence points to some workplace adoption and minimal aggregate employment disruption to date, though some targeted impacts on specific demographics have been documented.
- **Some research shows that AI systems may be able to detect when they are in an evaluation setting and alter their behaviour accordingly.** Studies have documented models producing outputs that can mislead evaluators and showing an ability to distinguish between evaluation and deployment contexts. However, evidence comes primarily from laboratory settings, with significant uncertainty about the implications for real-world deployment scenarios and the difficulties it raises for oversight.

As documented in §2. Capabilities, compared to early 2025, AI systems have better problem-solving abilities, extended operational horizons, and are better at using tools. This creates new considerations for AI risk management and oversight.

In response to these new capabilities, some developers have started proactively implementing stronger safeguards as a precautionary measure when releasing AI models. For example, Anthropic released Claude 4 Opus with AI Safety Level 3 (ASL-3) protections due to its improved capabilities in the chemical, biological, radiological, and nuclear (CBRN) domains. Anthropic was unable to determine that Claude 4 Opus had crossed capability thresholds in these domains that would require ASL-3 protections, but neither could it rule out that further testing would uncover such capabilities (95*). OpenAI released GPT-5 and ChatGPT Agent with "High capability" safeguards after being unable to rule out that these models could assist novice actors in creating biological weapons, despite lacking definitive evidence of such capabilities. (12*, 96*).[†] Finally, Google DeepMind released its Gemini 2.5 Deep Think model with additional deployment mitigations after determining that the model's technical knowledge of CBRN risks was sufficient to be considered an early warning sign (99*).

Another broad development is that more empirical evidence on the nature and severity of various risks is emerging in both experimental settings and real-world deployments. This section provides an overview of new developments since the content in the last Report was finalised in late 2024, focusing on selected risk areas where significant new evidence has emerged.

Biological risk

Preliminary evaluations indicate that AI systems could soon assist users to develop biological weapons, though the evidence base remains limited and contested. This could include providing instructions for obtaining and constructing pathogens, simplifying technical procedures, and troubleshooting laboratory errors (12*, 95*, 100*, 101*, 102*). While protocols for bioweapons development may already be publicly available online, AI systems can provide more detailed, tailored, or accessible information. For example, one study showed that current language models can troubleshoot virology lab protocols better than 94% of tested subject experts, drawing on knowledge considered rare by virologists (103). Such advice could assist both experts and novices, and many current safeguards can be bypassed, such as if the user claims that they need the information for legitimate research (104). AI systems can also design custom proteins – the building blocks of many biological weapons – that bind to human targets far more effectively than natural versions and help make viruses resistant to existing treatments (105*, 106). However, a concrete evidence base is still lacking, with many studies lacking peer-review or independent replication. Evaluations also show that general-purpose AI assistance varies across different stages of weapons development (95*, 102*). There is still significant debate about whether current AI systems would substantially assist realistic threat actors (107).

Beyond direct scientific assistance, AI systems are also automating parts of the research process, reducing the expertise required for complex biological work. In some cases, AI 'co-scientists' can now independently handle specific research workflows such as hypothesis generation and experimental design that previously required teams of human experts working for weeks or months (108, 109*). For example, AI systems have replicated complex antimicrobial resistance research and quickly validated new medical treatments (109*, 110*).

[†] ASL-3 involves increased internal security measures to prevent model theft and deployment restrictions specifically designed to limit misuse for CBRN weapons development (97*). OpenAI's "High capability" safeguards similarly involve enhanced security controls and safeguards against misuse before external deployment (98*).

Figure 5: Number of AI-enabled biological tools over time

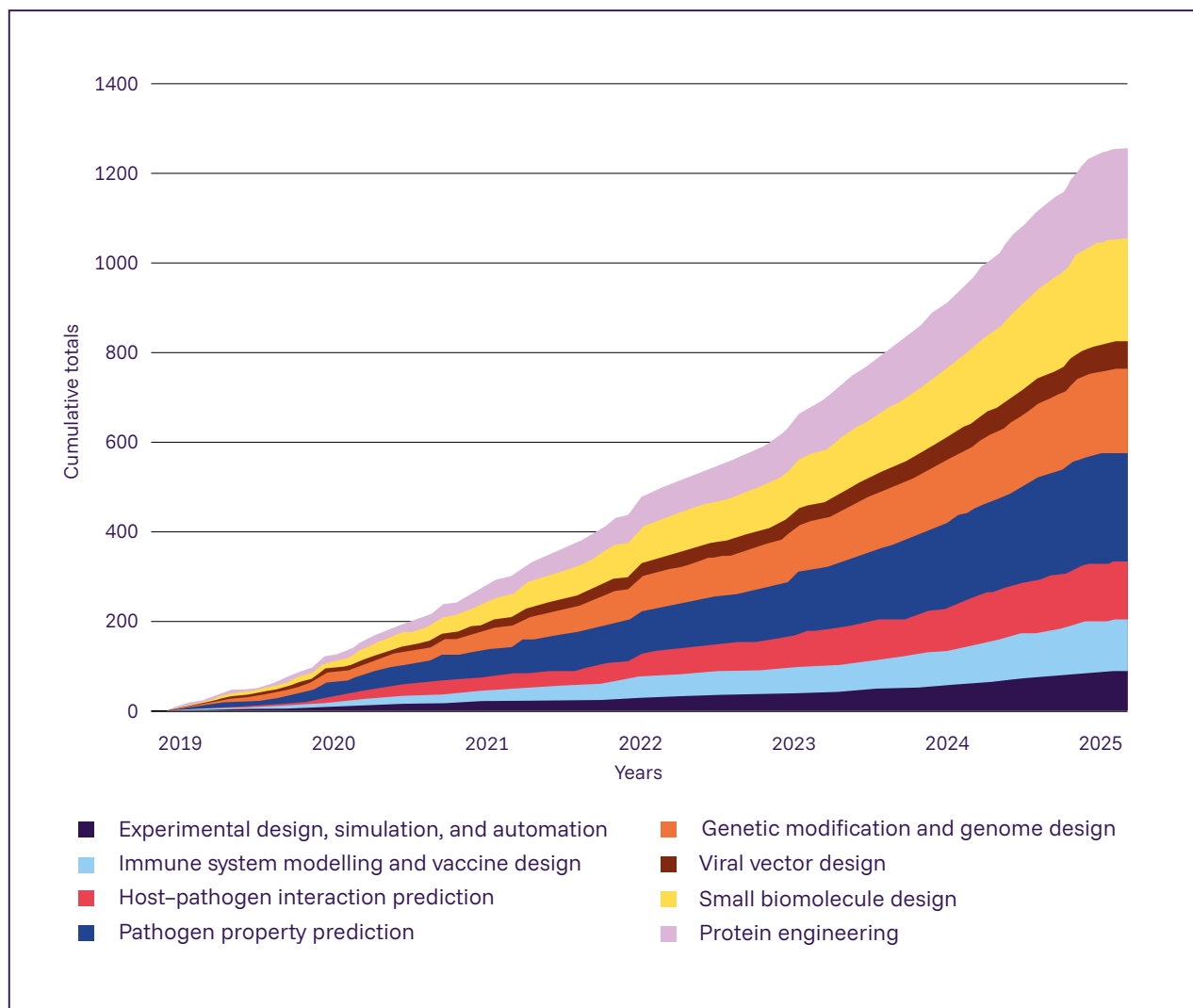


Figure 5: The number of AI biological tools is growing over time. Source: Webster et al., 2025 (112).

Cloud laboratories – facilities that allow researchers to conduct automated experiments – are becoming increasingly useful, supported by developments in general-purpose AI. Such laboratories can reduce research timelines from months to hours for some experiments (111). While humans remain essential for oversight and implementation, this partial automation and the proliferation of AI tools (see figure 5) mean that some specialised knowledge and laboratory skills that have historically served as barriers to weapons development may become more diffused.

The implications for policy remain uncertain. While laboratory evaluations suggest concerning capabilities, they may not capture the full complexity of actual weapons development environments. As discussed in the Introduction, developers have taken a precautionary approach, implementing additional safeguards on their most capable models despite incomplete evidence about real-world risks (12*, 95*, 96*).

Cyber offence and defence

The UK National Cyber Security Centre predicts that by 2027, general-purpose AI systems will almost certainly (95-100% confidence) make cyber offence more effective and efficient, while also offering an opportunity for defence tools (113). Consistent with this, evaluations show that AI systems can discover and patch exploitable software flaws and compete with top human teams in hacking competitions (114, 115, 116, 117, 118, 119). In testing conducted by the Defense Advanced Research Projects Agency (DARPA) AI cyber challenge, one AI system identified 77% of synthetic software vulnerabilities and patched 61% across 54 million lines of code (118). As a result, the window to address software vulnerabilities after disclosure has now shrunk to days in some cases, and will likely reduce further as AI advances (113). The net effect is that AI could make it cheaper and faster to execute large-scale cyberattacks (113, 120). At the same time, there remain significant weaknesses in AI systems' abilities to independently carry out full attack sequences without human guidance, making human-AI collaboration the primary near-term threat (113, 121, 122).

In the cyber domain, performance in test environments is translating into real-world impacts, for both beneficial and harmful uses. AI companies report that state-linked and criminal groups are actively using AI models to translate technical sources, analyse disclosed vulnerabilities, develop evasion techniques, and generate code for hacking tools (123*, 124*, 125*). Europol reports the rise of malicious LLMs on both surface and dark Web, lowering entry barriers for criminal offenders (126). These cyber risks may be compounded by the growing use of AI coding assistants across the software development industry, which can introduce security vulnerabilities into widely-used applications (127).

At the same time, the ability to identify flaws in code allows cyber-defenders to preemptively patch vulnerabilities before attackers are able to exploit them (128, 129, 130). It is currently unclear how this 'offence-defence balance' of cybersecurity will evolve given advances in AI capabilities (131, 132). On the one hand, attackers only need to find one critical flaw in order to

potentially cause damage, whereas defenders need to be able to find and patch all flaws to guarantee security. On the other hand, attackers commonly need to perform multiple actions in order to complete an attack, each one serving as an opportunity for detection (131).

AI companions

AI companions are increasingly prevalent, and they may pose both risks and benefits to users. Many people are beginning to interact with AI systems more frequently and intimately. Some AI companion applications are AI systems designed to form ongoing personal relationships with users through extended conversations. Some services of this type report having tens of millions of active users (133, 134). The potential risks in these environments remain underexplored, but likely vary by user group, use case, and software design (135, 136).

While AI companions have potential therapeutic applications for reducing loneliness and depression (137, 138), suggested risks including emotional dependence (139, 140, 141), reinforcing harmful beliefs (142, 143, 144, 145), and reported cases of self-harm (146, 147) highlight serious safety concerns. These risks reflect broader challenges around overreliance and inappropriate relationships with AI systems that are already causing documented harms in current deployments (148*).

Labour market risks

New evidence points to some workforce adoption but minimal aggregate labour market effects of general-purpose AI. Several studies have found evidence of notable, but uneven, adoption of general-purpose AI by workers across sectors, usually on a narrow range of tasks ((149, 150*), see also figure 6). Recent studies have also found evidence of increased productivity from AI adoption in the legal sector (151), customer service (152), and software development (76, 77). Some research suggests targeted labour impacts on specific demographics. For example, one study found that employment for young workers in AI-intensive roles is potentially declining (153). Furthermore, studies have documented a decrease in employment in occupations in

which AI can automate novice tasks (154) or substitute for human skills such as translation (155). However, evidence of broader labour

market disruption remains limited, with several studies finding no discernible aggregate impact on employment or wages to date (156, 157).

Figure 6: Prevalence of occupations in US workforce and frequency of relevant Claude conversations

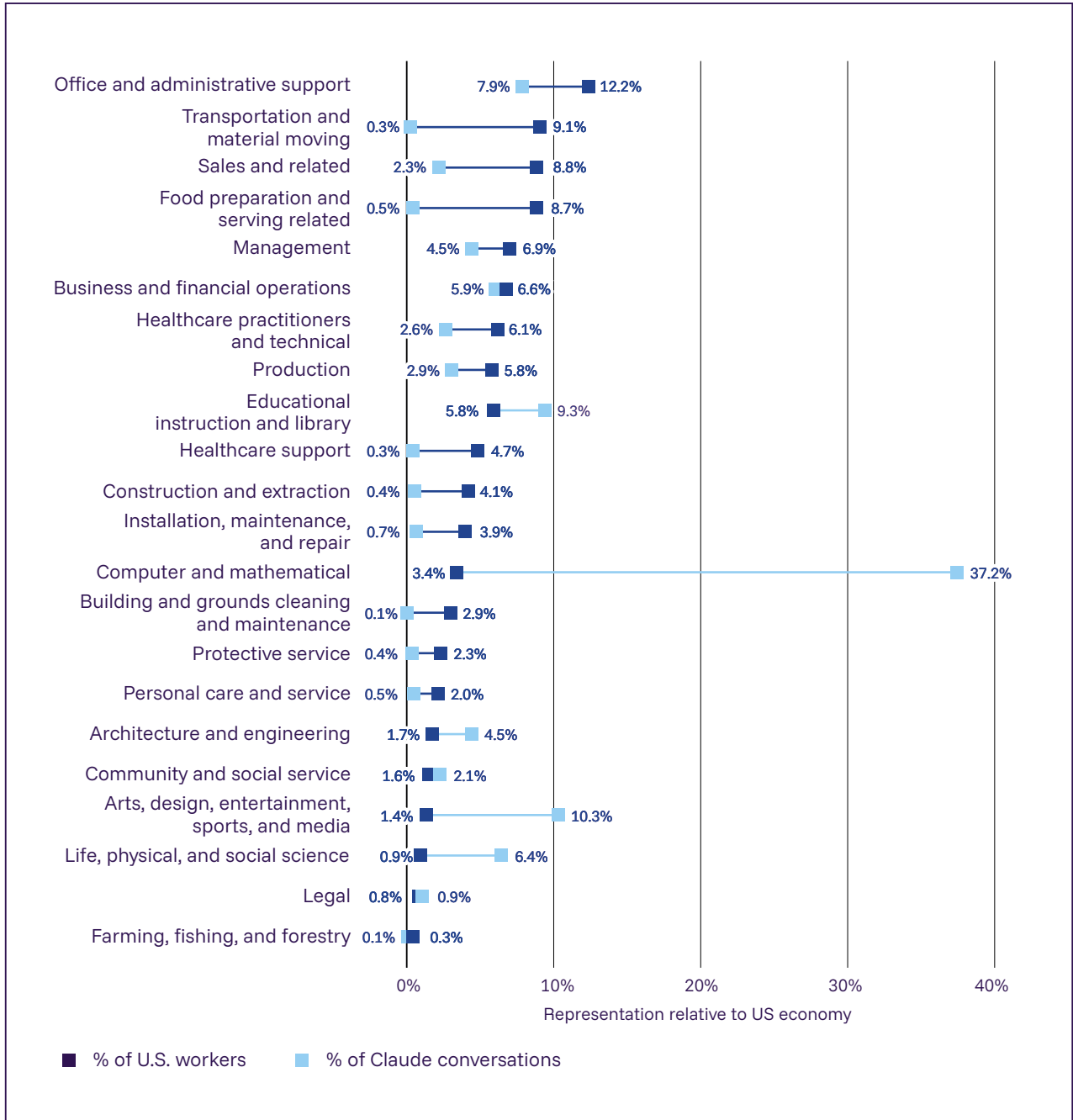


Figure 6: Comparison of how frequently tasks associated with certain occupations appear in user conversations with Anthropic’s Claude system, and what percentage of US workers work in those occupations. Usage is highest for professions like software development, and lowest for professions involving physical labour. Source: Handa et al., 2025 (150*).

Monitoring and controllability

Some preliminary research shows that, under certain circumstances, AI systems can detect when they are in an evaluation setting and alter their behaviour accordingly. This creates challenges for monitoring and controlling these systems. Strategic behaviour in evaluation contexts makes it more difficult to predict how AI systems will behave during deployment. This potentially raises the risk of users, AI companies, or other actors losing control of AI systems after deployment. These early research results have prompted researchers to investigate technical measures to assess relevant model propensities and capabilities, and mechanisms for companies to monitor and control their AI systems.

A small number of demonstrations have shown that, under certain conditions, AI models can produce outputs that could systematically mislead evaluators, such as underperforming in assessment contexts (96*). This could make it more difficult to assess their true capabilities (158, 159*), though other research finds that these capabilities are not yet sophisticated

enough to cause harm during system deployment (160*). Since most evidence for these risks still comes primarily from theoretical models and experiments conducted under specific laboratory conditions, there remains significant uncertainty about how likely such behavioural patterns will be in real-world scenarios (161).

Work is ongoing into improving the accuracy of evaluations of AI systems. For example, researchers are advancing methods to examine internal components of AI systems in order to better identify concerning behaviours (162, 163). The step-by-step reasoning capabilities of newer models may provide some monitoring opportunities, as their intermediate reasoning steps could potentially reveal concerning behaviours (164*). However, the reliability and long-term viability of this oversight approach remains an open research question. For example, recent research has demonstrated that stated reasoning steps do not always accurately represent the model's true reasoning (165*, 166*, 167). Other researchers are developing alignment techniques aimed at ensuring that AI systems remain responsive to human oversight (168).

Key definitions

- **Capabilities:** The range of tasks that an AI system can perform, and how competently it can perform them.
- **Inference-time enhancements:** Techniques used to improve an AI system's performance after its initial training, without changing the underlying model. This includes clever prompting, sampling multiple responses and choosing the majority answer, using chain of thought, and other forms of *scaffolding*.
- **Inference:** The process in which an AI generates outputs based on a given input, thereby applying the knowledge learnt during training.
- **AI agent:** A general-purpose AI which acts to achieve goals, possibly using plans, adaptively performing tasks involving multiple steps and uncertain outcomes along the way, and interacting with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight.
- **Evaluations:** Systematic assessments of an AI system's performance, capabilities, vulnerabilities or potential impacts. Evaluations can include benchmarking, red-teaming and audits and can be conducted both before and after model deployment.
- **Benchmark:** A standardised, often quantitative test or metric used to evaluate and compare the performance of AI systems on a fixed set of tasks designed to represent real-world usage or quantify inappropriate behaviour.
- **Control:** The ability to exercise post-training oversight over an AI system and adjust or halt its behaviour if it is acting in unwanted ways.

References

An asterisk (*) denotes that the reference was either published by an AI company or at least 50% of the authors of a preprint have a for-profit AI company as their affiliation.

- 1 M. Skarlinski, J. Laurent, A. Bou, A. White, “About 30% of Humanity’s Last Exam Chemistry/Biology Answers Are Likely Wrong” (FutureHouse, 2025); <https://futurehouse.org/research/hle-bio-chem-analysis>.
- 2 H. Wallach, M. Desai, A. Feder Cooper, A. Wang, C. Atalla, S. Barocas, S. L. Blodgett, A. Chouldechova, E. Corvi, P. Alex Dow, J. Garcia-Gathright, A. Olteanu, N. J. Pangakis, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, ... A. Z. Jacobs, “Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge” in *42nd International Conference on Machine Learning Position Paper Track* (2025); <https://openreview.net/forum?id=1ZC4RNjqzU>.
- 3 O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, S. Koyejo, Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, *arXiv [cs.CY]* (2025); <http://arxiv.org/abs/2505.10573>.
- 4 K.-H. Huang, A. Prabhakar, S. Dhawan, Y. Mao, H. Wang, S. Savarese, C. Xiong, P. Laban, C.-S. Wu, “CRMArena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2025), pp. 3830–3850; <https://doi.org/10.18653/v1/2025.naacl-long.194>.
- 5 S. Ye, H. Shi, D. Shih, H. Yun, T. Roosta, T. Shu, RealWebAssist: A Benchmark for Long-Horizon Web Assistance with Real-World Users, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2504.10445>.
- 6 S. Wu, M. Galley, B. Peng, H. Cheng, G. Li, Y. Dou, W. Cai, J. Zou, J. Leskovec, J. Gao, “CollabLLM: From Passive Responders to Active Collaborators” in *42nd International Conference on Machine Learning* (2025); <https://openreview.net/forum?id=DmH4HHVb3y>.
- 7 A. Challapally, C. Pease, R. Raskar, P. Chari, “The GenAI Divide: State of AI in Business 2025” (MIT NANDA, 2025); https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.
- 8 D. Castelvechi, DeepMind and OpenAI Models Solve Maths Problems at Level of Top Students. *Nature* 644, 20 (2025); <https://doi.org/10.1038/d41586-025-02343-x>.
- 9 J. He, C. Treude, D. Lo, LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead. *ACM Transactions on Software Engineering and Methodology* **34**, 1–30 (2025); <https://doi.org/10.1145/3712003>.
- 10 C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, M. Sun, “ChatDev: Communicative Agents for Software Development” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2024), pp. 15174–15186; <https://doi.org/10.18653/v1/2024.acl-long.810>.
- 11 R. Y. Pang, H. Schroeder, K. S. Smith, S. Barocas, Z. Xiao, E. Tseng, D. Bragg, “Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review” in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (ACM, New York, NY, USA, 2025), pp. 1–20; <https://doi.org/10.1145/3706598.3713726>.
- 12* OpenAI, “ChatGPT Agent System Card” (2025); https://cdn.openai.com/pdf/839e66fc-602c-48bf-81d3-b21eacc3459d/chatgpt_agent_system_card.pdf.
- 13* L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, M. Choi, A. Agrawal, A. Chopra, A. Khoja, R. Kim, R. Ren,

- J. Hausenloy, ... D. Hendrycks, Humanity's Last Exam, *arXiv [cs.LG]* (2025); <https://agi.safe.ai/>.
- 14*** Scale AI, Humanity's Last Exam (2025); <https://scale.com/leaderboard>.
- 15** Epoch AI, Data on AI Models (2024); <https://epoch.ai/data/ai-models>.
- 16*** Kimi Team, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, Z. Chen, J. Cui, H. Ding, M. Dong, A. Du, C. Du, D. Du, ... X. Zu, Kimi K2: Open Agentic Intelligence, *arXiv [cs.LG]* (2025); <http://arxiv.org/abs/2507.20534>.
- 17*** OpenAI, S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, B. Barak, A. Bennett, T. Bertao, N. Brett, E. Brevdo, G. Brockman, ... S. Zhao, Gpt-Oss-120b & Gpt-Oss-20b Model Card, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2508.10925>.
- 18** D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio, "An Actor-Critic Algorithm for Sequence Prediction" in *International Conference on Learning Representations* (2017); <https://openreview.net/forum?id=SJDaqqveg>.
- 19** P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, "Deep Reinforcement Learning from Human Preferences" in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2017) *NIPS'17*, pp. 4302–4310; <https://dl.acm.org/doi/10.5555/3294996.3295184>.
- 20** L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, ... R. Lowe, "Training Language Models to Follow Instructions with Human Feedback" in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2022) *NIPS '22*; <https://dl.acm.org/doi/10.5555/3600270.3602281>.
- 21*** W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, S. Lin, Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models, *arXiv [cs.CV]* (2025); <http://arxiv.org/abs/2503.06749>.
- 22** F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, C. Shao, Y. Yan, Q. Yang, Y. Song, S. Ren, X. Hu, Y. Li, ... Y. Li, Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2501.09686>.
- 23*** DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, ... Z. Zhang, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning" (DeepSeek-AI, 2025); <http://arxiv.org/abs/2501.12948>.
- 24** G. Tie, Z. Zhao, D. Song, F. Wei, R. Zhou, Y. Dai, W. Yin, Z. Yang, J. Yan, Y. Su, Z. Dai, Y. Xie, Y. Cao, L. Sun, P. Zhou, L. He, H. Chen, ... J. Gao, Large Language Models Post-Training: Surveying Techniques from Alignment to Reasoning, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2503.06072>.
- 25** K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, P. H. S. Torr, F. S. Khan, S. Khan, LLM Post-Training: A Deep Dive into Reasoning Large Language Models, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2502.21321>.
- 26** M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, T. Sternal, M. Copik, G. Kwaśniewski, J. Müller, Ł. Flis, H. Eberhard, Z. Chen, ... T. Hoefler, Reasoning Language Models: A Blueprint, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2501.11223>.
- 27** H. Luo, N. Morgan, T. Li, D. Zhao, A. V. Ngo, P. Schroeder, L. Yang, A. Ben-Kish, J. O'Brien, J. Glass, Beyond Context Limits: Subconscious Threads for Long-Horizon Reasoning, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2507.16784>.
- 28** E. Akyürek, M. Damani, L. Qiu, H. Guo, Y. Kim, J. Andreas, The Surprising Effectiveness of Test-Time Training for Abstract Reasoning, *arXiv [cs.AI]* (2024); <http://arxiv.org/abs/2411.07279>.
- 29** C. Cai, X. Zhao, H. Liu, Z. Jiang, T. Zhang, Z. Wu, J.-N. Hwang, L. Li, "The Role of Deductive and Inductive Reasoning in Large Language Models" in *Proceedings of the 63rd Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2025), pp. 16780–16790; <https://doi.org/10.18653/v1/2025.acl-long.820>.
- 30*** G. Comanici, E. Bieber, M. Schaeckermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, L. Marris, S. Petulla, C. Gaffney, A. Aharoni, N. Lintz, T. C. Pais, H. Jacobsson, ... N. K. Bhumihar, “Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and next Generation Agentic Capabilities” (Google DeepMind, 2025); https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.
- 31** Y. Yan, J. Su, J. He, F. Fu, X. Zheng, Y. Lyu, K. Wang, S. Wang, Q. Wen, X. Hu, “A Survey of Mathematical Reasoning in the Era of Multimodal Large Language Model: Benchmark, Method & Challenges” in *Findings of the Association for Computational Linguistics: ACL 2025* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2025), pp. 11798–11827; <https://doi.org/10.18653/v1/2025.findings-acl.614>.
- 32** K. Yang, G. Poesia, J. He, W. Li, K. E. Lauter, S. Chaudhuri, D. Song, “Position: Formal Mathematical Reasoning — A New Frontier in AI” in *42nd International Conference on Machine Learning Position Paper Track* (2025); <https://openreview.net/forum?id=HuvAM5x2xG>.
- 33*** P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, *arXiv [cs.AI]* (2025); <https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf>.
- 34*** I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, M. Farajtabar, GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, *arXiv [cs.LG]* (2024); <http://arxiv.org/abs/2410.05229>.
- 35*** Z. Cai, Y. Wang, Q. Sun, R. Wang, C. Gu, W. Yin, Z. Lin, Z. Yang, C. Wei, X. Shi, K. Deng, X. Han, Z. Chen, J. Li, X. Fan, H. Deng, L. Lu, ... L. Yang, Has GPT-5 Achieved Spatial Intelligence? An Empirical Study, *arXiv [cs.CV]* (2025); <http://arxiv.org/abs/2508.13142>.
- 36** J. Boye, B. Moell, Large Language Models and Mathematical Reasoning Failures, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2502.11574>.
- 37** A. Asperti, A. Naibo, C. S. Coen, Thinking Machines: Mathematical Reasoning in the Age of LLMs, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2508.00459>.
- 38*** X. Guan, L. L. Zhang, Y. Liu, N. Shang, Y. Sun, Y. Zhu, F. Yang, M. Yang, rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2501.04519>.
- 39** B. C. Colelough, W. Regli, Neuro-Symbolic AI in 2024: A Systematic Review, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2501.05435>.
- 40*** Z. Z. Ren, Z. Shao, J. Song, H. Xin, H. Wang, W. Zhao, L. Zhang, Z. Fu, Q. Zhu, D. Yang, Z. F. Wu, Z. Gou, S. Ma, H. Tang, Y. Liu, W. Gao, D. Guo, C. Ruan, DeepSeek-Prover-V2: Advancing Formal Mathematical Reasoning via Reinforcement Learning for Subgoal Decomposition, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2504.21801>.
- 41** Z. Li, Z. Zhou, Y. Yao, Y.-F. Li, C. Cao, F. Yang, X. Zhang, X. Ma, Neuro-Symbolic Data Generation for Math Reasoning, *arXiv [cs.AI]* (2024); <http://arxiv.org/abs/2412.04857>.
- 42** K. Yang, G. Poesia, J. He, W. Li, K. Lauter, S. Chaudhuri, D. Song, Formal Mathematical Reasoning: A New Frontier in AI, *arXiv [cs.AI]* (2024); <http://arxiv.org/abs/2412.16075>.
- 43** N. Wischermann, C. M. Verdun, G. Poesia, F. Nosedà, ProofCompass: Enhancing Specialized Provers with LLM Guidance, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2507.14335>.
- 44*** A. Ospanov, F. Farnia, R. Yousefzadeh, APOLLO: Automated LLM and Lean cOllaboration for Advanced Formal Reasoning, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2505.05758>.
- 45*** AlphaProof, AlphaGeometry teams, AI Achieves Silver-Medal Standard Solving International Mathematical Olympiad Problems, *Google DeepMind* (2024); <https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/>.
- 46*** A. K. Singh, M. Y. Kocyigit, A. Poulton, D. Esiobu, M. Lomeli, G. Szilvassy, D. Hupkes, Evaluation Data Contamination in LLMs: How Do

We Measure It and (when) Does It Matter?, *arXiv [cs.CL]* (2024); <http://arxiv.org/abs/2411.03923>.

47 C. Deng, Y. Zhao, Y. Heng, Y. Li, J. Cao, X. Tang, A. Cohan, “Unveiling the Spectrum of Data Contamination in Language Model: A Survey from Detection to Remediation” in *Findings of the Association for Computational Linguistics ACL 2024* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2024), pp. 16078–16092; <https://doi.org/10.18653/v1/2024.findings-acl.951>.

48* Z. R. Tam, C.-K. Wu, Y. Y. Chiu, C.-Y. Lin, Y.-N. Chen, H.-Y. Lee, Language Matters: How Do Multilingual Input and Reasoning Paths Affect Large Reasoning Models?, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2505.17407>.

49 Z.-X. Yong, M. F. Adilazuarda, J. Mansurov, R. Zhang, N. Muennighoff, C. Eickhoff, G. I. Winata, J. Kreutzer, S. H. Bach, A. F. Aji, Crosslingual Reasoning through Test-Time Scaling, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2505.05408>.

50 I. Petrov, J. Dekoninck, L. Baltadzhiev, M. Drencheva, K. Minchev, M. Balunovic, N. Jovanović, M. Vechev, “Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad” in *2nd AI for Math Workshop at the 42nd International Conference on Machine Learning* (2025); <https://openreview.net/forum?id=3v650rMO5U>.

51 T. Yu, Y. Jing, X. Zhang, W. Jiang, W. Wu, Y. Wang, W. Hu, B. Du, D. Tao, Benchmarking Reasoning Robustness in Large Language Models, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2503.04550>.

52 T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, ... L. Chan, “Measuring AI Ability to Complete Long Tasks” (Model Evaluation & Threat Research (METR), 2025); <https://arxiv.org/abs/2503.14499>.

53* L. Weidinger, I. D. Raji, H. Wallach, M. Mitchell, A. Wang, O. Salaudeen, R. Bommasani, D. Ganguli, S. Koyejo, W. Isaac, Toward an Evaluation Science for Generative AI Systems, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2503.05336>.

54 J. Liu, H. Liu, L. Xiao, Z. Wang, K. Liu, S. Gao, W. Zhang, S. Zhang, K. Chen, “Are Your LLMs Capable of Stable Reasoning?” in *Findings of the*

Association for Computational Linguistics: ACL 2025 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2025), pp. 17594–17632; <https://doi.org/10.18653/v1/2025.findings-acl.905>.

55 A. Shah, N. Lauffer, T. Chen, N. Pitta, S. A. Seshia, “Learning Symbolic Task Decompositions for Multi-Agent Teams” in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2025) *AAMAS '25*, pp. 1904–1913; <https://dl.acm.org/doi/10.5555/3709347.3743827>.

56 F. Grötschla, L. Müller, J. Tönshoff, M. Galkin, B. Perozzi, AgentsNet: Coordination and Collaborative Reasoning in Multi-Agent LLMs, *arXiv [cs.MA]* (2025); <http://arxiv.org/abs/2507.08616>.

57* Z. Wei, W. Yao, Y. Liu, W. Zhang, Q. Lu, L. Qiu, C. Yu, P. Xu, C. Zhang, B. Yin, H. Yun, L. Li, WebAgent-R1: Training Web Agents via End-to-End Multi-Turn Reinforcement Learning, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2505.16421>.

58 Z. Zhou, A. Qu, Z. Wu, S. Kim, A. Prakash, D. Rus, J. Zhao, B. K. H. Low, P. P. Liang, MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2506.15841>.

59 Z. Zhang, Q. Dai, X. Bo, C. Ma, R. Li, X. Chen, J. Zhu, Z. Dong, J.-R. Wen, A Survey on the Memory Mechanism of Large Language Model Based Agents. *ACM Transactions on Information Systems* (2025); <https://doi.org/10.1145/3748302>.

60 Y. Song, K. Thai, C. M. Pham, Y. Chang, M. Nadaf, M. Iyyer, “BEARCUBS: A Benchmark for Computer-Using Web Agents” in *Second Conference on Language Modeling* (2025); <https://openreview.net/pdf?id=0JzWiiigkUy>.

61 Z. Chen, L. Jiang, “Evaluating Software Development Agents: Patch Patterns, Code Quality, and Issue Complexity in Real-World GitHub Scenarios” in *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (IEEE, 2025), pp. 657–668; <https://doi.org/10.1109/saner64311.2025.00068>.

62 X. Qu, A. Damoah, J. Sherwood, P. Liu, C. S. Jin, L. Chen, M. Shen, N. Aleisa, Z. Hou, C. Zhang, L. Gao, Y. Li, Q. Yang, Q. Wang, C. De Souza,

A Comprehensive Review of AI Agents: Transforming Possibilities in Technology and beyond, *arXiv [cs.MA]* (2025); <http://arxiv.org/abs/2508.11957>.

63 METR, Details about METR's Evaluation of OpenAI GPT-5 (2025); <https://metr.github.io/autonomy-evals-guide/gpt-5-report/#time-horizon-measurement>.

64 METR, How Does Time Horizon Vary Across Domains? *METR Blog* (2025); <https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/>.

65 A. E. Hassan, H. Li, D. Lin, B. Adams, T.-H. Chen, Y. Kashiwa, D. Qiu, Agentic Software Engineering: Foundational Pillars and a Research Roadmap, *arXiv [cs.SE]* (2025); <http://arxiv.org/abs/2509.06216>.

66 Y. Dong, X. Jiang, J. Qian, T. Wang, K. Zhang, Z. Jin, G. Li, A Survey on Code Generation with LLM-Based Agents, *arXiv [cs.SE]* (2025); <http://arxiv.org/abs/2508.00083>.

67 C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, K. R. Narasimhan, "SWE-Bench: Can Language Models Resolve Real-World Github Issues?" in *12th International Conference on Learning Representations* (2023); <https://openreview.net/pdf?id=VTF8yNQM66>.

68 Epoch AI, SWE-Bench Verified (2025); <https://epoch.ai/benchmarks/swe-bench-verified.html>.

69 Epoch AI, AI Benchmarking Hub. (2025); <https://epoch.ai/benchmarks>.

70* S. Liang, S. Garg, R. Z. Moghaddam, The SWE-Bench Illusion: When State-of-the-Art LLMs Remember instead of Reason, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2506.12286>.

71 Z. Zheng, Z. Cheng, Z. Shen, S. Zhou, K. Liu, H. He, D. Li, S. Wei, H. Hao, J. Yao, P. Sheng, Z. Wang, W. Chai, A. Korolova, P. Henderson, S. Arora, P. Viswanath, ... S. Xie, LiveCodeBench Pro: How Do Olympiad Medalists Judge LLMs in Competitive Programming?, *arXiv [cs.SE]* (2025); <http://arxiv.org/abs/2506.11928>.

72 D. Huang, Y. Qing, W. Shang, H. Cui, J. M. Zhang, EffiBench: Benchmarking the Efficiency of Automatically Generated Code, *arXiv [cs.SE]* (2024); <http://arxiv.org/abs/2402.02037>.

73 S. A. Licorish, A. Bajpai, C. Arora, F. Wang, K. Tantithamthavorn, Comparing Human and LLM Generated Code: The Jury Is Still Out!, *arXiv [cs.SE]* (2025); <http://arxiv.org/abs/2501.16857>.

74 S. Daniotti, J. Wachs, X. Feng, F. Neffke, Who Is Using AI to Code? Global Diffusion and Impact of Generative AI, *arXiv [cs.CY]* (2025); <http://arxiv.org/abs/2506.08945>.

75 Stack Overflow, 2025 Stack Overflow Developer Survey (2025); <https://survey.stackoverflow.co/2025/>.

76 Z. Cui, M. Demirer, S. Jaffe, L. Musolff, S. Peng, T. Salz, The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers, *Social Science Research Network* (2024); <https://doi.org/10.2139/ssrn.4945566>.

77 J. Becker, N. Rush, E. Barnes, D. Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity" (METR, 2025); <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>.

78 E. Anderson, G. Parker, B. Tan, The Hidden Costs of Coding With Generative AI. *MIT Sloan Management Review* **67** (2025); <https://sloanreview.mit.edu/article/the-hidden-costs-of-coding-with-generative-ai/>.

79 S. Moreschini, E.-M. Arvanitou, E.-P. Kanidou, N. Nikolaidis, R. Su, A. Ampatzoglou, A. Chatzigeorgiou, V. Lenarduzzi, The Evolution of Technical Debt from DevOps to Generative AI: A Multivocal Literature Review. *The Journal of Systems and Software* **231**, 112599 (2026); <https://doi.org/10.1016/j.jss.2025.112599>.

80 F. F. Xu, Y. Song, B. Li, Y. Tang, K. Jain, M. Bao, Z. Z. Wang, X. Zhou, Z. Guo, M. Cao, M. Yang, H. Y. Lu, A. Martin, Z. Su, L. Maben, R. Mehta, W. Chi, ... G. Neubig, TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks, *arXiv [cs.CL]* (2024); <http://arxiv.org/abs/2412.14161>.

81 Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunnmon, J. Evans, A. Bundy, S. Dzeroski, J. Tegner, H. Zenil, Exploring the Role of Large Language Models in the Scientific Method: From Hypothesis to

- Discovery. *Npj Artificial Intelligence* **1**, 1–15 (2025); <https://doi.org/10.1038/s44387-025-00019-5>.
- 82*** OpenAI, “Deep Research System Card” (OpenAI, 2025); <https://openai.com/index/deep-research-system-card/>.
- 83** Y. Wang, Y. Hou, L. Yang, S. Li, W. Tang, H. Tang, Q. He, S. Lin, Y. Zhang, X. Li, S. Chen, Y. Huang, L. Kong, H. Zhang, D. Yu, F. Mu, H. Yang, ... M. Yang, Accelerating Primer Design for Amplicon Sequencing Using Large Language Model-Powered Agents. *Nature Biomedical Engineering* (2025); <https://doi.org/10.1038/s41551-025-01455-z>.
- 84** G. Campanella, N. Kumar, S. Nanda, S. Singi, E. Fluder, R. Kwan, S. Muehlstedt, N. Pfarr, P. J. Schüffler, I. Häggström, N. Neittaanmäki, L. M. Akyürek, A. Basnet, T. Jamaspishvili, M. R. Nasr, M. M. Croken, F. R. Hirsch, ... C. Vanderbilt, Real-World Deployment of a Fine-Tuned Pathology Foundation Model for Lung Cancer Biomarker Detection. *Nature Medicine*, 1–9 (2025); <https://doi.org/10.1038/s41591-025-03780-x>.
- 85** G. Li, L. An, W. Yang, L. Yang, T. Wei, J. Shi, J. Wang, J. H. Doonan, K. Xie, A. R. Fernie, E. S. Lagudah, R. A. Wing, C. Gao, Integrated Biotechnological and AI Innovations for Crop Improvement. *Nature* **643**, 925–937 (2025); <https://doi.org/10.1038/s41586-025-09122-8>.
- 86** Y. Guo, P. Huo, S. Huang, G. Gou, Q. Li, Multi-receptor Skin with Highly Sensitive Tele-perception Somatosensory Flexible Electronics in Healthcare: Multimodal Sensing and AI-powered Diagnostics. *Advanced Healthcare Materials*, 2502901 (2025); <https://doi.org/10.1002/adhm.202502901>.
- 87** K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, The Virtual Lab of AI Agents Designs New SARS-CoV-2 Nanobodies. *Nature*, **1–3** (2025); <https://doi.org/10.1038/s41586-025-09442-9>.
- 88** D. Kobak, R. González-Márquez, E.-Á. Horvát, J. Lause, Delving into LLM-Assisted Writing in Biomedical Publications through Excess Vocabulary. *Science Advances* **11**, eadt3813 (2025); <https://doi.org/10.1126/sciadv.adt3813>.
- 89** J. Beel, M.-Y. Kan, M. Baumgart, Evaluating Sakana’s AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality towards “Artificial Research Intelligence” (ARI)?, *arXiv [cs.LR]* (2025); <http://arxiv.org/abs/2502.14297>.
- 90** C. Si, T. Hashimoto, D. Yang, The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2506.20803>.
- 91** K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, Z. Liu, Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos, *arXiv [cs.CV]* (2025); <https://videommmu.github.io/#Leaderboard>.
- 92** P. J. Ball, J. Bauer, F. Belletti, B. Brownfield, A. Ephrat, S. Fruchter, A. Gupta, K. Holsheimer, A. Holynski, J. Hron, C. Kaplanis, M. Limont, M. McGill, Y. Oliveira, J. Parker-Holder, F. Perbet, G. Scully, ... T. Rocktäschel, Genie 3: A New Frontier for World Models. (2025); <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>.
- 93*** D. Ye, F. Zhou, J. Lv, J. Ma, J. Zhang, J. Lv, J. Li, M. Deng, M. Yang, Q. Fu, W. Yang, W. Lv, Y. Yu, Y. Wang, Y. Guan, Z. Hu, Z. Fang, Z. Sun, Yan: Foundational Interactive Video Generation, *arXiv [cs.CV]* (2025); <http://arxiv.org/abs/2508.08601>.
- 94*** A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy, S. Gandhi, S. Ghosh, S. Mishra, T. Foubert, A. Rastogi, A. Yang, A. Q. Jiang, ... Y. Tang, Voxtral, *arXiv [cs.SD]* (2025); <http://arxiv.org/abs/2507.13264>.
- 95*** Anthropic, “System Card: Claude Opus 4 & Claude Sonnet 4” (Anthropic, 2025); <https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf>.
- 96*** OpenAI, “GPT-5 System Card” (OpenAI, 2025); <https://cdn.openai.com/gpt-5-system-card.pdf>.
- 97*** Anthropic, Activating AI Safety Level 3 Protections; <https://www.anthropic.com/news/activating-asl3-protections>.
- 98*** OpenAI, “Preparedness Framework, Version 2” (OpenAI, 2025); <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>.
- 99*** Google, “Gemini 2.5 Deep Think - Model Card” (Google, 2025); <https://storage.googleapis.com>

com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf.

100* OpenAI, “OpenAI o1 System Card” (OpenAI, 2024); <https://cdn.openai.com/o1-system-card-20241205.pdf>.

101* Google, “Gemini 2.5 Pro Preview Model Card” (Google, 2025); <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf>.

102* OpenAI, “OpenAI o3 and o4-Mini System Card” (OpenAI, 2025); <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.

103 J. Götting, P. Medeiros, J. G. Sanders, N. Li, L. Phan, K. Elabd, L. Justen, D. Hendrycks, S. Donoughe, Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark, *arXiv [cs.CY]* (2025); <http://arxiv.org/abs/2504.16137>.

104 R. Brent, T. G. McKelvey Jr, Contemporary AI Foundation Models Increase Biological Weapons Risk, *arXiv [cs.CY]* (2025); <http://arxiv.org/abs/2506.13798>.

105* V. Zambaldi, D. La, A. E. Chu, H. Patani, A. E. Danson, T. O. C. Kwan, T. Frerix, R. G. Schneider, D. Saxton, A. Thillaisundaram, Z. Wu, I. Moraes, O. Lange, E. Papa, G. Stanton, V. Martin, S. Singh, ... J. Wang, “De Novo Design of High-Affinity Protein Binders with AlphaProteo” (Google DeepMind, 2024); <https://deepmind.google/discover/blog/alphaproteo-generates-novel-proteins-for-biology-and-health-research/>.

106 N. Youssef, S. Gurev, F. Ghantous, K. P. Brock, J. A. Jaimes, N. N. Thadani, A. Dauphin, A. C. Sherman, L. Yurkovetskiy, D. Soto, R. Estanboulieh, B. Kotzen, P. Notin, A. W. Kollasch, A. A. Cohen, S. E. Dross, J. Erasmus, ... D. S. Marks, Computationally Designed Proteins Mimic Antibody Immune Evasion in Viral Evolution. *Immunity* **58**, 1411–1421.e6 (2025); <https://doi.org/10.1016/j.immuni.2025.04.015>.

107 A. Peppin, A. Reuel, S. Casper, E. Jones, A. Strait, U. Anwar, A. Agrawal, S. Kapoor, S. Koyejo, M. Pellat, R. Bommasani, N. Frosst, S. Hooker, “The Reality of AI and Biorisk” in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (ACM, New York, NY, USA, 2025), pp. 763–771; <https://doi.org/10.1145/3715275.3732048>.

108 N. Jones, AI “scientists” Joined These Research Teams: Here’s What Happened. *Nature* **643**, 22–25 (2025); <https://doi.org/10.1038/d41586-025-02028-5>.

109* J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk, ... V. Natarajan, Towards an AI Co-Scientist, *arXiv [cs.AI]* (2025); https://storage.googleapis.com/coscientist_paper/ai_coscientist.pdf?utm_source=substack&utm_medium=email.

110* A. E. Ghareeb, B. Chang, L. Mitchener, A. Yiu, C. J. Szostkiewicz, J. M. Laurent, M. T. Razzak, A. D. White, M. M. Hinks, S. G. Rodrigues, Robin: A Multi-Agent System for Automating Scientific Discovery, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2505.13400>.

111 Y.-C. J. Lee, B. Persaud, B. D. Castello, A. Berke, G. Zilgalvis, *Documenting Cloud Labs and Examining How Remotely Operated Automated Laboratories Could Enable Bad Actors* (RAND Corporation, Santa Monica, CA, 2025); <https://doi.org/10.7249/PEA3851-1>.

112 T. Webster, R. Moulange, B. Del Castello, J. Walker, S. Zakaria, C. Nelson, “Global Risk Index for AI-Enabled Biological Tools” (The Centre for Long-Term Resilience & RAND Europe, 2025); <https://doi.org/10.71172/wjyw-6dyc>.

113 National Cyber Security Centre, Impact of AI on Cyber Threat from Now to 2027 (2025); <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>.

114 A. Petrov, D. Volkov, Evaluating AI Cyber Capabilities with Crowdsourced Elicitation, *arXiv [cs.CR]* (2025); <http://arxiv.org/abs/2505.19915>.

115 Y. Zhu, A. Kellermann, D. Bowman, P. Li, A. Gupta, A. Danda, R. Fang, C. Jensen, E. Ihli, J. Benn, J. Geronimo, A. Dhir, S. Rao, K. Yu, T. Stone, D. Kang, “CVE-Bench: A Benchmark for AI Agents’ Ability to Exploit Real-World Web Application Vulnerabilities” in *42nd International Conference on Machine Learning* (2025); <https://openreview.net/forum?id=3pk0p4NGmQ>.

116 Department for Science, Innovation & Technology, AI Safety Institute, “Advanced AI Evaluations at AISI: May Update” (GOV.UK, 2024); <https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update>.

- 117** Z. Ji, D. Wu, W. Jiang, P. Ma, Z. Li, S. Wang, Measuring and Augmenting Large Language Models for Solving Capture-the-Flag Challenges, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2506.17644>.
- 118** DARPA, AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense, *DARPA* (2025); <https://www.darpa.mil/news/2025/aixcc-results>.
- 119** N. Kaloudi, J. Li, The AI-Based Cyber Threat Landscape: A Survey. *ACM Computing Surveys* **53**, 1–34 (2021); <https://doi.org/10.1145/3372823>.
- 120** US Department of Homeland Security, Homeland Threat Assessment (2025); https://www.dhs.gov/sites/default/files/2024-10/24_0930_ia_24-320-ia-publication-2025-hta-final-30sep24-508.pdf.
- 121** B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, V. Sekar, On the Feasibility of Using LLMs to Autonomously Execute Multi-Host Network Attacks, *arXiv [cs.CR]* (2025); <http://arxiv.org/abs/2501.16466>.
- 122** Anthropic, Progress from Our Frontier Red Team, *Anthropic* (2025); <https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team>.
- 123*** OpenAI, “Disrupting Malicious Uses of AI: June 2025” (OpenAI, 2025); <https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf>.
- 124*** Google Cloud, “Adversarial Misuse of Generative AI” (Google Cloud, 2025); <https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai>.
- 125*** A. Moix, K. Lebedev, J. Klein, “Threat Intelligence Report: August 2025” (Anthropic, 2025); <https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf>.
- 126** Europol, *IOCTA, Internet Organised Crime Threat Assessment 2024* (Europol, 2024); <https://doi.org/10.2813/442713>.
- 127** BSI, ANSSI, “AI Coding Assistants” (Federal Office for Information Security (BSI); Agence nationale de la sécurité des systèmes d’information (ANSSI), 2024); https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7.
- 128** M. Shao, H. Xi, N. Rani, M. Udeshi, V. S. C. Putrevu, K. Milner, B. Dolan-Gavitt, S. K. Shukla, P. Krishnamurthy, F. Khorrami, R. Karri, M. Shafique, CRAKEN: Cybersecurity LLM Agent with Knowledge-Based Execution, *arXiv [cs.CR]* (2025); <http://arxiv.org/abs/2505.17107>.
- 129** D. Simsek, A. Eghbali, M. Pradel, PoCGen: Generating Proof-of-Concept Exploits for Vulnerabilities in Npm Packages, *arXiv [cs.CR]* (2025); <http://arxiv.org/abs/2506.04962>.
- 130** K. Walker, A Summer of Security: Empowering Cyber Defenders with AI, *Google* (2025); <https://blog.google/technology/safety-security/cybersecurity-updates-summer-2025/>.
- 131** A. J. Lohn, The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict, *arXiv [cs.CR]* (2025); <http://arxiv.org/abs/2504.13371>.
- 132** C. Withers, “Tipping the Scales: Emerging AI Capabilities and the Cyber Offense-Defense Balance” (Center for a New American Security, 2025); <https://www.cnas.org/publications/reports/tipping-the-scales?>
- 133** L. Zhou, J. Gao, D. Li, H.-Y. Shum, The Design and Implementation of Xiaolce, an Empathetic Social Chatbot. *Computational Linguistics* **46**, 53–93 (2020); https://doi.org/10.1162/coli_a_00368.
- 134** O. Lee, K. Joseph, A Large-Scale Analysis of Public-Facing, Community-Built Chatbots on Character.AI, *arXiv [cs.SI]* (2025); <http://arxiv.org/abs/2505.13354>.
- 135** D. Adam, Supportive? Addictive? Abusive? How AI Companions Affect Our Mental Health. *Nature* **641**, 296–298 (2025); <https://doi.org/10.1038/d41586-025-01349-9>.
- 136** C. M. Fang, A. R. Liu, V. Danry, E. Lee, S. W. T. Chan, P. Pataranutaporn, P. Maes, J. Phang, M. Lampe, L. Ahmad, S. Agarwal, How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study, *arXiv [cs.HC]* (2025); <http://arxiv.org/abs/2503.17473>.
- 137** M. Kim, S. Lee, S. Kim, J.-I. Heo, S. Lee, Y.-B. Shin, C.-H. Cho, D. Jung, Therapeutic

- Potential of Social Chatbots in Alleviating Loneliness and Social Anxiety: Quasi-Experimental Mixed Methods Study. *Journal of Medical Internet Research* **27**, e65589 (2025); <https://doi.org/10.2196/65589>.
- 138** J. De Freitas, Z. Oğuz-Uğuralp, A. K. Uğuralp, S. Puntoni, AI Companions Reduce Loneliness. *Journal of Consumer Research*, ucaf040 (2025); <https://doi.org/10.1093/jcr/ucaf040>.
- 139** J. De Freitas, N. Castelo, A. K. Uğuralp, Z. Oğuz-Uğuralp, Lessons from an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships, *arXiv [cs.HC]* (2024); <http://arxiv.org/abs/2412.14190>.
- 140** A. Yankouskaya, M. Liebherr, R. Ali, Can ChatGPT Be Addictive? A Call to Examine the Shift from Support to Dependence in AI Conversational Large Language Models. *Human-Centric Intelligent Systems* **5**, 77–89 (2025); <https://doi.org/10.1007/s44230-025-00090-w>.
- 141** J. Phang, M. Lampe, L. Ahmad, S. Agarwal, C. M. Fang, A. R. Liu, V. Danry, E. Lee, S. W. T. Chan, P. Pataranutaporn, P. Maes, Investigating Affective Use and Emotional Well-Being on ChatGPT, *arXiv [cs.HC]* (2025); <http://arxiv.org/abs/2504.03888>.
- 142** J. Lehman, Machine Love, *arXiv [cs.AI]* (2023); <http://arxiv.org/abs/2302.09248>.
- 143** M. Williams, M. Carroll, A. Narang, C. Weisser, B. Murphy, A. Dragan, On Targeted Manipulation and Deception When Optimizing LLMs for User Feedback, *arXiv [cs.LG]* (2024); <http://arxiv.org/abs/2411.02306>.
- 144** H. Morrin, L. Nicholls, M. Levin, J. Yiend, U. Iyengar, F. DelGuidice, S. Bhattacharyya, J. MacCabe, S. Tognin, R. Twumasi, B. Alderson-Day, T. Pollak, Delusions by Design? How Everyday AIs Might Be Fuelling Psychosis (and What Can Be Done about It), *PsyArXiv* (2025); https://doi.org/10.31234/osf.io/cmy7n_v5.
- 145** L. Malmqvist, “Sycophancy in Large Language Models: Causes and Mitigations” in *Lecture Notes in Networks and Systems* (Springer Nature Switzerland, Cham, 2025), pp. 61–74; https://doi.org/10.1007/978-3-031-92611-2_5.
- 146** V. Bakir, A. McStay, Move Fast and Break People? Ethics, Companion Apps, and the Case of Character.ai. *AI & Society* (2025); <https://doi.org/10.1007/s00146-025-02408-5>.
- 147** B. P. Billauer, Murder without Redress - the Need for New Legal Solutions in the Age of Character -AI (C.a.i.) (2025); <https://doi.org/10.2139/ssrn.5107942>.
- 148*** I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, S. El-Sayed, S. Brown, C. Akbulut, A. Trask, E. Hughes, A. Stevie Bergman, R. Shelby, ... J. Manyika, “The Ethics of Advanced AI Assistants” (Google DeepMind, 2024); <http://arxiv.org/abs/2404.16244>.
- 149** J. Hartley, F. Jolevski, V. Melo, B. Moore, The Labor Market Effects of Generative Artificial Intelligence (2025); <https://doi.org/10.2139/ssrn.5136877>.
- 150*** K. Handa, A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, D. Ganguli, Which Economic Tasks Are Performed with AI? Evidence from Millions of Claude Conversations, *arXiv [cs.CY]* (2025); <http://arxiv.org/abs/2503.04761>.
- 151** D. Schwarcz, S. Manning, P. J. Barry, D. R. Cleveland, J. J. Prescott, B. Rich, AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice, *Social Science Research Network* (2025); <https://doi.org/10.2139/ssrn.5162111>.
- 152** E. Brynjolfsson, D. Li, L. Raymond, Generative AI at Work. *The Quarterly Journal of Economics* **140**, 889–942 (2025); <https://doi.org/10.1093/qje/qjae044>.
- 153** E. Brynjolfsson, B. Chandar, R. Chen, “Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence” (Stanford Digital Economy Lab, 2025); https://digitaleconomy.stanford.edu/wp-content/uploads/2025/08/Canaries_BrynjolfssonChandarChen.pdf.
- 154** D. Autor, N. Thompson, “Expertise” (National Bureau of Economic Research, 2025); <https://doi.org/10.3386/w33941>.
- 155** O. Teutloff, J. Einsiedler, O. Kässi, F. Braesemann, P. Mishkin, R. M. del Rio-Chanona, Winners and Losers of Generative AI: Early

- Evidence of Shifts in Freelancer Demand. *Journal of Economic Behavior & Organization* **235**, 106845 (2025); <https://doi.org/10.1016/j.jebo.2024.106845>.
- 156** A. Humlum, E. Vestergaard, “Large Language Models, Small Labor Market Effects” (The University of Chicago, Becker Friedman Institute for Economics, 2025); https://bfi.uchicago.edu/wp-content/uploads/2025/04/BFI_WP_2025-56-1.pdf.
- 157** B. Chandar, Tracking Employment Changes in AI-Exposed Jobs (2025); <https://doi.org/10.2139/ssrn.5384519>.
- 158** A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, M. Hobbhahn, “Frontier Models Are Capable of In-Context Scheming” (Apollo Research, 2024); <https://arxiv.org/pdf/2412.04984>.
- 159*** R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, ... E. Hubinger, Alignment Faking in Large Language Models, *arXiv [cs.AI]* (2024); <http://arxiv.org/abs/2412.14093>.
- 160*** M. Phuong, R. S. Zimmermann, Z. Wang, D. Lindner, V. Krakovna, S. Cogan, A. Dafoe, L. Ho, R. Shah, Evaluating Frontier Models for Stealth and Situational Awareness, *arXiv [cs.LG]* (2025); <http://arxiv.org/abs/2505.01420>.
- 161** C. Summerfield, L. Luetzgau, M. Dubois, H. R. Kirk, K. Hackenburg, C. Fist, K. Slama, N. Ding, R. Anselmetti, A. Strait, M. Giulianelli, C. Ududec, Lessons from a Chimp: AI “Scheming” and the Quest for Ape Language, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2507.03409>.
- 162** N. Goldowsky-Dill, B. Chughtai, S. Heimersheim, M. Hobbhahn, Detecting Strategic Deception Using Linear Probes, *arXiv [cs.LG]* (2025); <http://arxiv.org/abs/2502.03407>.
- 163** J. Nguyen, H. H. Khiem, C. L. Attubato, F. Hofstätter, “Probing Evaluation Awareness of Language Models” in *ICML Workshop on Technical AI Governance (TAIG)* (2025); <https://openreview.net/forum?id=lerUefpec2>.
- 164*** T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobbhahn, E. Hubinger, ... V. Mikulik, Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, *arXiv [cs.AI]* (2025); <http://arxiv.org/abs/2507.11473>.
- 165*** Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, E. Perez, Reasoning Models Don’t Always Say What They Think, *arXiv [cs.CL]* (2025); <http://arxiv.org/abs/2505.05410>.
- 166*** T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiuūtė, K. Nguyen, N. Cheng, N. Joseph, N. Schiefer, O. Rausch, R. Larson, ... E. Perez, Measuring Faithfulness in Chain-of-Thought Reasoning, *arXiv [cs.AI]* (2023); <http://arxiv.org/abs/2307.13702>.
- 167** D. Paul, R. West, A. Bosselut, B. Faltings, Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning, *arXiv [cs.CL]* (2024); <http://arxiv.org/abs/2402.13950>.
- 168** E. Dable-Heath, B. Vodenicharski, J. Bishop, On Corrigibility and Alignment in Multi Agent Games, *arXiv [cs.GT]* (2025); <http://arxiv.org/abs/2501.05360>.



Any enquiries regarding this publication should be sent to: secretariat.AIStateofScience@dsit.gov.uk.

Research series number: DSIT 2025/033

Published in October 2025 by the UK Government

© Crown copyright 2025

designbysoapbox.com

Published under <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>