Understanding AI Trajectories: Mapping the Limitations of Current AI Systems

Max Heitmann[1], Ture Hinrichsen[2], David Africa[3], and Jonas Sandbrink[4]

[1,2,3,4]UK AI Security Institute

As of September 2025, AI systems lack the general cognitive capabilities that could enable large-scale automation of cognitive labour. The limitations of today's systems include performance limitations on certain types of tasks, insufficient reliability, insufficient adaptability, and low capacity for original insights. Much of the progress on these limitations to date has been driven by scaling up the computational resources used to train and run AI systems. This trend is expected to continue, but there is uncertainty over how much future progress will require fundamental research innovation versus being driven primarily by further compute scaling. Through literature review, interviews with internal and external experts, and an internal workshop with the UK AISI Research Unit, we have identified eight indicators of progress against the limitations of current AI systems, and gathered perspectives on how these limitations are most likely to be overcome. Each indicator is presented alongside an initial appraisal of the evidence for progress and suggestions for additional pieces of evidence that would bear on it.

1 Introduction

1.1 Motivation

Some experts believe that scaling up current methods for building and training AI systems will produce systems capable of automating most cognitive labour. Such systems, if assimilated into the labour force, would have significant implications for our economies and wider society, ranging from increased productivity to widespread displacement of human workers. As of September 2025, frontier AI systems fall short of this level of capability in several respects. This paper categorises the limitations of today's AI systems, and investigates the state of the evidence for progress against them. It should be emphasised that we are not making a normative claim about whether the automation of most cognitive labour would be desirable, but rather examining technical limitations of current systems across relevant capabilities.

Continued scaling (increasing computational resources) could drive progress against many of the limitations discussed in this report. Reporting from Epoch AI in August 2024 suggests that current rates of compute scaling can plausibly be sustained for another five years through to 2030, give or take a year [1]. Beyond that, however, the challenges associated with further scaling of computing infrastructure become much more significant. For example, the researchers find that, on current trends, the largest individual frontier training runs in 2030 will require 4–16 GW of power, which is enough to power millions of US homes [2]. This suggests that, to the extent that limitations are not overcome by around 2030, more of the burden of progress may need to shift to research innovation on underlying architectures, algorithms, and scaffolding techniques rather than scaling. Nevertheless, research progress is also likely to occur before 2030, and scaling may continue (albeit at a slower pace) beyond 2030. Accordingly, this report considers compute scaling alongside research progress as potential solutions for overcoming limitations.

1.2 Scope of this report

We focus on the limitations of current systems (as of 09/25) insofar as those limitations pose a bottleneck for reaching AI systems that can automate most cognitive labour (i.e., acting as a "drop-in remote worker"). In this paper, we refer to this sort of AI system as "AGI", in line with how the term

is used in OpenAI's charter from 2018 ('highly autonomous systems that outperform humans at most economically valuable work') [3] and Google DeepMind's "Levels of AGI" framework paper ('an AI system that is at least as capable as a human at most cognitive tasks') [4].[1] This definition is related to, but distinct from, the characterisation of AGI as an AI system that poses catastrophic risk through loss of control (for a discussion of loss of control and advancements towards control-undermining capabilities, see the International AI Safety Report 2025 [6]). While the report is framed around this milestone, we recognise that many impacts of highly capable AI systems may emerge gradually, with no requirement to 'cross AGI' for large economic or security effects to materialise.

Out of scope: hardware progress. While we do include compute scaling as a technical approach to overcoming limitations, this report does not discuss the hardware enablers of further compute scaling such as the number and efficiency of the GPUs used for training AI systems.

Out of scope: automation of AI R&D. Several experts emphasized that a key milestone on the path to AGI is the automation of AI R&D. Their argument is that once AI R&D is automated, the effective size of the research workforce increases dramatically, accelerating the rate of progress, and consequently we should expect the remaining research bottlenecks to fall rapidly (compute remains a bottleneck).[2] In this scenario, tracking particular research bottlenecks would be significantly less useful than tracking indicators for AI R&D automation. However, not all experts agreed that the automation of AI R&D is likely to precede the creation of full AGI by a notable amount. Therefore, while this trajectory should be taken seriously, we consider this out of scope of this paper.

Out of scope: robotics. This report focusses on bottlenecks to developing AI systems that can automate all cognitive labour, and excludes physical labour.

1.3 Methodology

This paper is based on a literature review, interviews with researchers and other experts, and an internal workshop with the UK AISI Research Unit. Interviews were conducted both internally and externally. Within AISI, we interviewed several researchers (Joseph Bloom, Michael Schmatz, and Jacob Pfau), as well as AISI's Chief Scientist, Geoffrey Irving. Outside of AISI, we interviewed Daniel Kokotajlo (AI Futures Project), Ryan Greenblatt (Redwood Research), Benjamin Todd (80,000 Hours), Jacob Hilton (Alignment Research Center), Ajeya Cotra (Open Philanthropy), Rohin Shah (Google DeepMind), James Wilken-Smith (independent researcher), Alex Lawsen (Open Philanthropy), and researchers from the LLM-agent safety research group "Aether". We also received written comments from Arvind Narayanan (Princeton University) on a late-stage draft. We note that many of these experts are highly concerned about AGI risk in the short- to medium-term, and this report may not reflect perspectives that are more sceptical about deep learning and LLMs as a viable paradigm for AGI. Quality assurance was carried out through peer review by external experts and workshop participants, as well as review by AISI's chief scientist Geoffrey Irving (who is also listed above).

2 Limitations of current AI systems

This report is structured around four different categories of limitations of current AI systems. These are performance limitations on certain types of tasks, such as hard-to-verify tasks and long tasks (subsection 2.1); reliability limitations, such as difficulties with avoiding errors and hallucinations (subsection 2.2); adaptability limitations, such as limitations on the ability of AI systems to adapt to the deployment environment, and to continue learning after deployment (subsection 2.3); and limitations in AI systems' capacity for original insights of genuine scientific value (subsection 2.4).

Within these four categories, we have identified eight concrete limitations. Progress or stagnation on any of these limitations can serve as indicators for different possible AI trajectories. For each limitation, we assess (i) to what extent overcoming the limitation is necessary for achieving AGI, and (ii) to what extent the limitation is on track for being solved. Our findings are summarised in Table 1.

[1] While common, this definition is controversial. For instance, Arvind Narayanan and Sayash Kapoor have argued that whether an AI system can automate most cognitive labour has as much to do with the state of the world (e.g., the extent to which we choose to integrate AI into our organizations and institutions) as it does with the properties of the AI system itself. Additionally, the usefulness of this definition could be limited by the fact that we may not be able to definitively assess whether we have achieved AGI by this standard prior to the widespread diffusion of AI [5].

[2] For a discussion of this trajectory, see Forethought's report on the software intelligence explosion [7]. For an example of what this might look like if it happens in 2027, see the AI 2027 scenario forecast [8].

Indicator	To what extent is this needed for AGI?	To what extent is this on track for being solved?
Performance on tasks that are hard to verify (subsubsection 2.1.1) *Description*: Hard-to-verify tasks are those for which performance cannot be cheaply and accurately assessed at scale.	Much economically valuable work consists of hard-to-verify tasks. Examples include strategic decision-making, policymaking, and research & discovery.	There is evidence of modest progress on broader verification (e.g., for mathematical proofs), but experts disagreed about the evidence for strong generalisation, especially across subject domains.
Performance on long tasks (subsubsection 2.1.2) *Description*: Long tasks are tasks that take human workers a long time to complete, often involving many serial steps.	Much economically valuable work consists of long tasks. At least some of these tasks are likely intrinsically long, in the sense that they cannot be effectively broken up.	Benchmarks show sustained exponential progress on time-horizons for AI systems. Further, trends for 50% and 80% success rate obey similar doubling times.
Performance in complex environments (subsubsection 2.1.3) *Description*: Real-world environments tend to be larger and more unpredictable than the training environments for AI systems.	Better performance in complex environments is needed for AGI, but the technical challenges involved may overlap with indicators 2.1.1 and 2.1.2.	Evidence suggests that AI systems struggle in large, high-variability environments, and there is a gap between benchmarks and real-world utility. However, there is also limited evidence of progress.
Sufficiently low error rate (subsubsection 2.2.1) *Description*: Errors include simple mistakes, such as transcription errors and arithmetical errors, as well as logically invalid reasoning steps.	Certain deployment contexts have low error tolerance. However, the bar is not zero: humans also are not perfectly reliable.	Various benchmarks provide evidence for a decline in intrinsic error rates. There is also evidence for emergent error correction in frontier reasoning models, although this is currently liable to robustness failures.
Meta-awareness and calibration (subsubsection 2.2.2) *Description*: Meta-awareness is the ability to know what one knows and how one knows it, including adjusting confidence to the probability that a claim is true ("confidence calibration").	Confidence calibration is necessary for some real-world jobs, such as high-stakes decision making and forecasting. Again, perfect calibration is not the target, since humans are not perfectly calibrated.	Despite the verbal behaviour of certain commercial chatbots, the evidence indicates that frontier language models already demonstrate substantial capability in this area.
Adaptability to the deployment environment (subsubsection 2.3.1) *Description*: Deployment adaptability is the effective application of learned skills appropriate to the local work context.	Adaptability is needed to the extent that real-world jobs cannot be decomposed into discrete, well-defined tasks, which may characterise most jobs.	Adaptability is progressing on multiple fronts: context length and utilisation are improving rapidly, PEFT methods enable efficient weight adaptation, and specialised architectures for efficient long-context memory are in development.
Continual learning post-deployment (subsubsection 2.3.2) *Description*: Continual learning is the ability of an AI model to learn from a changing stream of data while retaining competence on what it learned before.	Strong "out-of-the-box" task performance could make up for a lack of continual learning in a static world, but this is less plausible when the demands of cognitive labour change over time.	Evidence shows that frontier LLMs have made significant progress on catastrophic forgetting, allowing for continual learning in principle. But sample efficiency and implementation remain major challenges.
Original insights (subsubsection 2.4.1) *Description*: Original insights are novel ideas of genuine scientific value.	While substantial automation may be possible with only moderately creative AI systems, certain jobs plausibly require genuine original insights (including AI R&D).	Despite modest recent successes with multi-agent systems, notable examples of AI systems having substantial original insights are rare.

Table 1: An overview of current AI system limitations and the progress towards overcoming them.

2.1 Performance limitations on certain types of tasks

Today's AI systems already perform at human or superhuman level on many tasks, but they still struggle on important categories of tasks for automating cognitive labour. Tasks where AI systems already excel are often short-form and self-contained, such as writing executive summaries, coding short computer programs, or solving mathematical problems with known solutions. Tasks where today's AI systems still struggle to match human performance include tasks where strong performance is hard to verify, long tasks, and tasks that take place in large, high-variability environments.

2.1.1 Indicator: Performance on tasks that are hard to verify

Hard-to-verify tasks are those for which performance cannot be cheaply and accurately assessed at scale. Current AI systems perform best on tasks where performance can be accurately assessed at scale, particularly in so-called "verifiable domains." Two notable examples are mathematical problem-solving tasks with a numerical right answer, and self-contained programming tasks with well-defined tests (like writing code that compiles). In both these cases, it is easy to check whether the model submitted a correct solution, and indeed verification can be automated. These properties enable large-scale training on the task. But not all tasks have these training-friendly properties.

Strong performance on hard-to-verify tasks is necessary for achieving AGI. A substantial proportion of economically valuable work involves tasks that do not have accurate, scalable verification procedures. Examples include tasks like strategic decision-making (success takes a long time to manifest and attribution is complex), policymaking (success depends on the counterfactual, which is unobservable), and tasks requiring aesthetic judgement (which are hard to rate objectively). According to our definition, AGI is AI that can automate most cognitive labour, so this includes strong performance on these tasks.

There are two broad categories of technical solutions: either (1) accurate, scalable performance assessment methods on a broader range of tasks, or (2) sufficient transfer of performance from easy-to-verify tasks to hard-to-verify tasks. While there is considerable evidence of progress on broader verification, experts disagreed about the state of the evidence for effective task transfer. We consider these in turn.

1. **Broader verification.** Many tasks cannot be easily and accurately verified at scale (e.g., by comparing the model's answer to a known ground truth). Using human graders is not very scalable, since the human capacity to grade model outputs is limited and cannot match the very large number of training samples needed for frontier training runs in 2025.[3] Another widely applied technique is to use AI graders [9]. Existing approaches include training a separate AI model specifically for the purpose of grading outputs ("reward models") or using the primary model to assess its own outputs for tasks where verification is systematically easier than solution generation. Approaches that use AI graders can be applied recursively, so that as the primary model improves with training, so too does the quality of the AI grader's assessments.

   The recent strong performance of OpenAI and Google DeepMind models at this year's International Mathematics Olympiad (IMO) [10] is evidence for success at broader verification, since IMO solutions take the form of mathematical arguments and natural-language proofs rather than easily verifiable numerical answers. Both OpenAI and Google DeepMind have suggested that they achieved their success by training their models on the task of writing mathematical proofs, rather than relying on task transfer from training on more easily verified tasks.

2. **Task transfer.** Instead of developing assessment methods for hard-to-verify tasks, an alternative approach is to modify the training algorithms in ways that improve the generalisation of the learned model behaviour. If successful, this strategy could facilitate performance improvements on hard-to-verify tasks merely through training on easily verifiable tasks. One expert suggested that better task transfer might also be achieved holding training algorithms fixed, by scaling the quantity and diversity of training environments. Then, even without training explicitly on hard-to-verify tasks, AI systems might pick up a variety of instrumentally useful skills that can be applied to such tasks (with a small amount of elicitation).

[3] For example, the training of DeepSeek-R1, released in January 2025, required the verification of 600k reasoning samples. DeepSeek used verifiable rewards, so this was straightforward; using human graders instead would've required significantly greater expense. And more recent frontier reasoning models very likely used an even greater number of data samples for RL training.

The key research frontier is better generalisation from training using reinforcement learning (RL). The idea is that, to a greater or lesser extent, RL methods teach reusable skills that can be composed together for good performance on categories of tasks beyond those seen during training. For example, there is evidence that reusable reasoning skills such as problem decomposition, backtracking, and error correction can emerge organically though outcome-based RL training on auto-gradable reasoning tasks alone.[4] This suggests that generalisation is certainly possible, but some experts expressed doubts over strong forms of RL generalisation, such as generalisation across subject domains.

Additional evidence that could be used to track progress against this limitation.

- Performance trends on capability evaluations that include hard-to-verify tasks. Notably, we already see some improvement on hard-to-verify tasks, so a useful indicator would be to track whether performance on hard-to-verify tasks is improving along similar trendlines but offset in time, or whether it is following entirely different (slower) trendlines.[5]
- Officially adjudicated wins on competitions involving problems that cannot be automatically graded, such as the International Maths Olympiad or various essay competitions.

2.1.2 Indicator: Performance on tasks that take humans a long time

Current AI systems have an advantage over humans on short tasks, but skilled humans are stronger on longer tasks.[6] This is because performance deteriorates much faster with task length for AI systems than for human workers. Experts proposed different theories for the underlying cause of this deterioration in task performance. According to one theory, the core problem is that AI models struggle with error correction, and so the probability of failure at the task compounds with each independent step, thus increasing exponentially with task length [14]. Another theory is that AI models struggle to remain coherently goal-directed over a long time, gradually drifting away from their original objective.

It is highly likely that better task performance on long tasks is necessary to achieve AGI. There is uncertainty over what fraction of economically valuable work can be decomposed into short tasks. However, it is plausible that at least some work involves tasks lasting days or weeks, rather than mere minutes or hours. For example, METR estimates that AI R&D automation would require task completion time horizons of at least a week [15]. Other examples might include litigating a complex court case, or directing a feature film from pre-production through to post-release.

The METR time-horizon benchmark [13] shows sustained exponential progress on time-horizons for AI systems, with a doubling time of approximately seven months at both 50% and 80% success rate.[7] This trend may be supported into the future by a number of technical approaches, including (1) scaling up existing methods, (2) improving the sample efficiency of RL algorithms, and (3) scaffolding-based techniques for adding executive function.

1. **Scaling up existing methods.** A relatively theory-agnostic approach to overcoming this limitation is to directly train AI models on longer tasks using outcome-based RL, without any substantial innovation in technical methods (e.g., algorithms or scaffolding). Additionally, imitation learning from human data can also be scaled up substantially, provided developers can collect enough data on human action sequences.

   Collecting the data to scale up imitation learning is costly but likely feasible: for example, major AI labs could take screen recordings of their employees while they work. But imitation learning

[4] See, for example, the DeepSeek "Aha!" moment [11]. DeepSeek do not disclose details of the datasets used to train R1-Zero, but they confirm that they took an all-RL approach with auto-gradable reasoning problems only—likely some mix of maths, logic, and LeetCode problems. In another result, researchers managed to elicit generalisable reasoning skills through training only on a single example [12].

[5] Related to this, the research non-profit METR has introduced "messiness factors" for different tasks and has begun comparing performance trends on messy vs. non-messy tasks; see Kwa et al. [13].

[6] METR: Measuring AI Ability to Complete Long Tasks [13]. Note that METR only measure AI task completion time horizons on software engineering tasks (of various types). We do not have enough data to judge what percentage of real-world long-horizon tasks are represented in this analysis, and it is possible that in some other domains AI systems still underperform human workers even on short tasks.

[7] METR's time-horizon benchmark measures the success rate of AI systems on a range of software and reasoning tasks against the time humans take to complete these tasks; the task set comprises a mix of RE-Bench [16] (for ML research engineering) and HCAST [17] (software engineering and general reasoning).

would need to be used in conjunction with RL to attain superhuman performance, and several experts mentioned the technical difficulties involved in directly training AI models on long tasks with RL. These include that training is less efficient due to sparse reward, constructing high-quality RL environments for long tasks is expensive and complicated, and the training time can get very long due to the need for serial long task completion. Despite the difficulties, developers are plausibly prepared to meet the costs of training on long tasks to some degree.

2. **Improving RL algorithms.** RL training is generally less sample efficient on long tasks, owing to the sparsity of the reward signal. Training can be made more sample efficient by improving credit assignment algorithms (methods for determining which past actions were responsible for observed outcomes so the agent can update its policy accordingly), or via direct evaluation of the reasoning process ("process rewards").

   The existing literature suggests that process rewards can often improve sample efficiency, especially on long multi-step problems. For example, Google DeepMind found a 5–6x gain in sample efficiency using dense rewards from "process advantage verifiers" over outcome rewards [18]; see also the PRIME algorithm for recent success with process rewards [19]. However, the advantage may not be universal, and this has not yet been studied systematically across many domains. And even if there is improvement, it is a further question whether the level of improvement is sufficient to enable training on very long tasks in practice.

3. **Scaffolding-based techniques for error correction and other aspects of executive function.** To the extent that the error correction theory or the goal drift theory capture salient aspects of the challenge, a direct solution that circumvents the needs for training on long tasks would be to put error- or goal drift-correction mechanisms into the model's thinking by hand, for example prompting the model every few steps to check for errors before continuing. More sophisticated proposals for using agent scaffolds to implement executive function have been suggested in informal commentary; see for example "internal independent review" [20].

   The existing literature paints a mixed picture for the potential of scaffolding techniques to add executive function. Scaffolding for self-reflection and iterative refinement, as implemented in Reflexion [21] and Self-Refine [22], seem to reliably boost error correction on several benchmarks. But the gains are modest and may be limited to tasks that can use reliable external feedback, such as tests and verifiers [23].

Additional evidence that could be used to track progress against this limitation.

- Continued progress on METR's time-horizon benchmark. To complement this, it would be instructive to track time-horizons on tasks beyond software engineering, to ensure that the trend is representative.
- Evidence of strong or improving beyond-episode generalisation—that is, the ability of AI systems to perform well on tasks with time-horizons longer than the training episode used for RL. Such generalisation could reduce the need for RL training on long tasks.
- Error correction and other "executive function" skills (e.g., deliberative techniques to stay on track for achieving the top-level objective) showing up in the chain-of-thought traces of reasoning models, and evidence that these tactics lead to higher success rates on long tasks (e.g., testing recovery from errors or distractions inserted into the chain-of-thought).

2.1.3 Indicator: Performance in complex environments

In addition to the above, experts communicated a general feeling that AI systems may struggle to perform well on certain real-world tasks because these take place in more complex environments. Real-world environments tend to be larger and more unpredictable than the environments used for RL training of AI systems, and strong performance in such environments requires a variety of skills which AI systems currently struggle with, such as effective prioritisation, coordinating with other agents, and dealing with novel situations.

Better performance in more complex environments is needed for AGI. Many jobs confront the human worker with a large range of possible situations (including novel / unprecedented situations),

in which they can take a large range of possible actions (including the full range of computer-use actions as well as physical actions and interactions with other agents). This contrasts with the more limited environments in which AI systems have so far been trained (e.g., the AI system can output tokens or use a limited number of tools). This discrepancy may partly explain the gap between benchmark performance and real-world usefulness of AI systems, since operating effectively in the real world requires dealing with a much larger, more unpredictable environment than the relatively well-circumscribed settings typical of most capability benchmarks.

There is evidence that AI systems do indeed struggle in these more complex environments, but there are also signs of progress. For example, METR conducted a randomised controlled trial on the impact of early-2025 AI on the productivity of experienced software developers, with the surprising finding that AI usage overall increased completion time for software development tasks [24]. This highlights a gap between benchmark performance and real-world utility, which might partly be explained by the additional complexity of operating in real-world environments. Supporting this, there is a growing body of anecdotal evidence around the difficulty that AI agents seem to encounter when forced to operate in complex real-world environments. Examples include Project Vend [25], in which Anthropic placed Claude 3.7 in charge of operating one of its office vending machines (it did not turn a profit), and AI Village [26], in which a teams of AI agents are given challenges such as raising money for charity. Additionally, METR's time-horizon study found that AI systems perform worse on "messier" tasks than would be predicted from task length alone, and several of METR's "messiness factors" relate to complex environments.[8] Importantly, METR also found that the rate of improvement on messy tasks is the same as for non-messy tasks, only offset in time, suggesting that we might see considerable progress on this limitation alongside progress on time horizons. However, since this result is taken from only one study, further evidence is required to make this judgement confidently.

Additional evidence that could be used to track progress against this limitation.

- In-the-wild evaluations placing AI systems in complex real-world test environments, such as Project Vend [25] and AI Village [26].
- Adoption rates in a wide variety of real-world environments and the effectiveness of AI in those environments. Concretely, one reviewer proposed tracking deployments of AI agents that (i) are general-purpose systems, (ii) operate with minimal supervision, and (iii) handle tasks with a high cost of errors.

2.2 Reliability limitations

Even if developers overcome the task performance limitations discussed in subsection 2.1, insufficient reliability may still block broad deployment of AI systems in certain contexts. Reliability limitations come in two main types: errors, where the model makes a mistake; and miscalibration, where the model does not match its confidence to the probability that it has produced the right answer.

2.2.1 Indicator: Error rates

Transformer-based AI systems sometimes make errors, such as arithmetical errors, transcription errors, and invalid reasoning steps. Some level of error is inevitable since transformer models do not compute with strict logical rules but instead work by generating probabilistically likely tokens. However, it is important to remember that humans also make errors, so the key question is not whether AI systems can be made entirely error-free but whether AI systems can attain sufficient reliability for a wide variety of real-world use cases. A closely related phenomenon is hallucinations, where AI systems present false, ungrounded answers as established facts. Hallucinations can be viewed as a consequence of imperfect accuracy in conjunction with the economic incentives of developers: AI models are trained to output an answer to every query, never withholding judgement, irrespective of whether they are capable of answering the question correctly [27].

Insufficient reliability in the form of high error rates could present an obstacle to real-world utility for two reasons. First, certain high-stakes deployment contexts (e.g., in defence and security systems,

[8] Measuring AI Ability to Complete Long Tasks [13], section 6.2. The messiness factors are listed in Tables 10 and 11, and relevant factors include *real life source*, *real-time coordination*, and *novel situation*.

or in safety-critical infrastructure) may have a very low error tolerance. If error rates cannot be driven down sufficiently, this could block deployment in these roles. Second, higher error rates may be associated with lower performance on long tasks. This is especially true in the absence of robust error correction mechanisms: in this case, high error rates equate to a rapidly compounding probability of failure on tasks with many independent sequential steps.

Error rates can be reduced in three ways: (1) by reducing the "instrinsic" rate at which AI systems make errors, (2) by improving error correction, or (3) by exploring alternative architectures that include a formal reasoning engine. Various benchmarks provide evidence for a decline in intrinsic error rates. The evidence for systematic emergent error correction is currently weaker, though there are signs that this could soon change.

1. **Intrinsic error rates.** Since high error rates reduce an AI model's task performance, training the model to improve at tasks using RL can be expected to drive down error rates. Conversely, better performance on tasks—as witnessed by trends in accuracy on standard performance benchmarks like GSM8K [28], MMLU [29], and GPQA [30]—can be taken as evidence for reduced error rates. Additionally, on the METR time horizon benchmark [13], time-horizons for 50% success rate and for 80% success rate follow the same exponential trend, the only difference being that 80% reliability lags 50% by a little over a year—this also suggests that intrinsic error rates have been consistently decreasing as capabilities improve.

   One expert emphasised that the crucial question is one of timing: errors need to be driven down sufficiently (i.e., below the tolerance threshold for broad deployment in various domains) before the challenges associated with further compute infrastructure scaling become much more significant.

2. **Error correction.** RL training reduces error rates both by reducing the intrinsic frequency with which the model makes errors, and by teaching the model how to correct its errors when it makes them. Organic error correction has been observed to occur in the chains-of-thought of current generation reasoning models. However, there are open questions about whether this can be made robust and systematic; see, for example, the problem of "late-stage fragility" in CoT reasoning [31], wherein early-stage errors can often trigger the reasoning model's latent self-correction mechanisms, whereas later errors are empirically harder for the model to identify and correct. Currently, the functional form of the decline in success rate with increasing task length for AI systems is still roughly exponential, in interesting contrast to the human case [14]. This provides some evidence for the constant hazard rate model of AI task performance time-horizons (i.e., a constant hazard rate leads to a compounding probability of failure at the task with increasing task length), which is indirect evidence that emergent error correction is not yet robustly implemented in reasoning models.

   While we emphasise that error correction is an emergent ability of AI models trained using RL, and indeed one which can be expected to improve with further scaling, developers can also supplement the foundation model's emergent error correction abilities with explicit mechanisms introduced into the system scaffold. For example, agent scaffolds could prompt the base model to check for errors every few steps during long serial reasoning tasks.

3. **Incorporate a formal reasoning engine.** Developers could explore hybrid "neuro-symbolic" architectures composed of a neural network working in conjunction with a formal reasoning engine (i.e., a program that executes reasoning steps according to strict logical rules). A number of powerful hybrid neural network / formal reasoning systems have been developed; a recent example is Google DeepMind's AlphaGeometry [32], which combined a neural language model with a symbolic deduction engine to achieve state-of-the-art performance on the International Maths Olympiad in 2024.[9] While these systems often achieve state-of-the-art performance in their domain of specialism (usually logic or mathematics) to our knowledge none of these systems rival the general competence of pure transformer models at today's frontier.

[9]Note: this is not the same DeepMind system that achieved an IMO gold medal in 2025, which was a pure-LLM reasoning model; for a discussion of the difference, see DeepMind's 2025 announcement [10].

Additional evidence that could be used to track progress against this limitation.

- Intrinsic error rates. Continued progress on benchmarks involving long serial reasoning problems, such as RE-Bench [16] and HCAST [17], especially when observed without evidence of systematic error correction. Progress on hallucinations can be monitored through benchmarks such as HHEM [33], HalluEval [34], and HalluLens [35].
- Error correction:
  - Progress on "late-stage fragility" [31] and other failures of robustness of the latent error correction abilities of LLMs. Models could be tested on data-sets cherry-picked for vulnerability to late-stage failures.
  - Empirical changes in the speed of performance decline in task completion success rates with increasing task length (i.e., observed functional form deviates from exponential).
- Formal reasoning. Improvements in general performance for AI models that employ a neuro-symbolic architecture. A key milestone would be if a major AI lab releases a state-of-the-art neuro-symbolic model such as AlphaGeometry, but which has frontier-level performance simultaneously across a range of domains (not just mathematics or logic).

2.2.2 Indicator: Meta-awareness and confidence calibration

Meta-awareness is the ability to know what one knows and how one knows it. This includes confidence calibration: the ability to accurately adjust one's confidence to the probability that one's own answer is correct. This is a distinct aspect of reliability from error rates: holding error rates fixed, an AI system that is aware when it is likely to be making errors is, in this sense, more reliable than an AI system that is not. Superficially, hallucinations seem to be an extreme example of a failure of confidence calibration, since AI models typically present their hallucinated answers with the same (high) level of confidence as they would for established facts, despite these answers lacking any evidential basis. However, we must be careful to disentangle genuine deficits of calibration from the verbal behaviour of commercial AI models, since commercial incentives may support training AI models to hallucinate confidently. This is because hallucinations allow AI models to produce interesting answers to a much broader range of queries. Moreover, confidence and authoritative tone are aspects of model personality that users might prefer.

Meta-awareness and confidence calibration are indispensable for many real-world jobs, such as jobs that involve high-stakes decision-making, placing bets, or forecasting. However, it is important to recognise that seemingly mis-calibrated outputs happen for a mixture of different reasons, and we must be careful to distinguish genuine failures of meta-awareness from other possible causes. For example, models may have been trained to adopt a sycophantic personality, which might involve confidently validating the user's beliefs irrespective of their evidential basis.[10] Additionally, there is documented evidence of deceptive outputs in certain settings;[11] if deceptive outputs are more prevalent, this can also generate a misleading appearance of mis-calibration. For our purposes, we set these other phenomena aside: a sycophantic personality or a propensity to deceive isn't a barrier to AGI in our sense, since these are not indicative of any lack of capabilities (though they could still pose a barrier to deployment for safety reasons). However, insofar as some cases of miscalibration are genuine, it seems plausible that modest progress against this limitation is needed for AGI.

Despite the superficial impression of poor meta-awareness, the evidence indicates that frontier language models already demonstrate substantial capability in this area (and have for some time). For example, already several years ago Anthropic found that language models exhibited strong levels of confidence calibration on diverse multiple-choice questions [38]. More recent work has explored other aspects of meta-awareness, such as behavioural self-awareness, with similar results [39]. Perhaps the primary data point against language model calibration is the behaviour of commercial chatbots, but we believe this can mostly be accounted for by reference to other causes.

[10] A particularly notable incident of this occurred with GPT-4o, in which OpenAI over-engaged certain aspects of likeable model personality [36].

[11] A striking example of this was observed already back in 2023, when researchers at METR (then, ARC) documented an instance of GPT-4 deliberately misleading a human TaskRabbit worker in order to solve a CAPTCHA [37].

Additional evidence that could be used to track progress against this limitation.

- Difference between stated confidence and actual correctness shrinks across domains. Metrics include benchmarks that directly assess language model calibration (such as GRACE [40]), or progress on tasks that require strong confidence calibration, such as forecasting [41].

2.3 Adaptability limitations

In addition to task performance and reliability, AI models may be limited by their adaptability. Here we take "adaptability" to subsume two main abilities. First, there is the adaptability of an instance of an AI model to the details of its local deployment environment. Second, there is the continual improvement of the AI model itself based on feedback it receives during real-world deployment. These abilities may be inter-substitutable to some degree: a deficit in continual learning may be compensated for by superior local adaptability, and vice versa.

2.3.1 Indicator: Adaptability to the deployment environment

In addition to possessing a large repertoire of learned skills, AI systems may need to have some ability to adapt to their local work environment to apply those skills most effectively. An analogy to the human case is instructive: when a human worker starts a new role, there is an initial period of onboarding and acclimatisation during which the human starts out relatively inept, but then rapidly improves at their job. This can be partly explained by the fact that new skills are being exercised, and performance is improving through practice of those skills. But most of the improvement comes from digesting lots of information about the organisation and its codes and practices, the specific role within the organisation, any materials that their predecessor left them, and so on. In other words, even while the skill profile of the employee remains largely fixed, their ability to meet the demands of the role improves rapidly as they absorb context from the work environment. Similarly, we might think that AI systems also need the ability to absorb a large amount of local context in order to function effectively in a real-world role.

Better adaptability to the deployment environment might not be strictly necessary for achieving AGI, but it might be one of the easiest pathways. Experts disagreed about its importance. One crux concerns the degree to which real-world jobs can be decomposed into discrete, well-defined tasks. Intuitively, to the extent that real-world jobs can be decomposed, strong task performance on its own may be sufficient to produce a high-value automated employee. Conversely, to the extent that real-world jobs cannot be easily decomposed, adaptability to local context might play a larger role in bridging the gap between strong performance on isolated tasks and the automation of full jobs. Given the already impressive performance of frontier models on many tasks required for the automation of cognitive labour, it is somewhat surprising that widespread automation of the labour market is not yet feasible. Better adaptability to local context could be a crucial missing piece.

Transformer-model outputs are conditioned by two things: the weights of the transformer model, and the tokens in the context window.[12] Thus, technical approaches to improving adaptability largely revolve around strategies for making better use of AI models' context windows, as well as methods for more efficient downstream fine-tuning. We go into more detail on these approaches below.

1. **Context length and utilisation.** Typical context window sizes for present generation transformers are hundreds of thousands of tokens in length, or roughly the equivalent of a dozen books' worth of text tokens (Google's Gemini 2.5 Pro currently leads on context window size, at 1 million tokens in length). Some of the experts maintained that the effective capacity of human short-term memory is considerably larger (which is more plausible when considering audio-visual information alongside text) and that this could partly explain why humans have the edge over AI systems when it comes to adaptability.

   Both context length and its effective utilization have been improving rapidly since mid-2023 [42]. However, expanding context windows in order to boost adaptability faces a number of technical challenges. For one, very large context models are more expensive to run and less efficient to

[12]The "context window" refers to the amount of input text, images, or audio tokens that a transformer model can process in a single forward pass through the network.

train. There is also evidence that the attention mechanism in transformers struggles over very long context windows; see, for instance, the "lost in the middle" problem [43] and the RULER benchmark [44]. Additionally, current models sometimes struggle to learn from their mistakes even when they are clearly visible in the context window—a failure of "in-context learning" (see section 2.3.2). For example, until quite recently reasoning models were still frequently losing at Tic-Tac-Toe the exact same way twice in a row, despite having access to the previous game's moves and results visible in their context window.[13] Weak in-context learning limits the usefulness of further expanding context windows as a method for improving adaptability.

2. **Improving context management.** An alternative to expanding the size of the context window is to better curate the content that fits into a context window of limited size. Human workers perform tasks by actively gathering relevant pieces of context (the problem being solved, past attempts at solving it, advice and suggestions from other people etc.) into short-term memory. By contrast, context management for AI systems is often fairly rudimentary, for example just using the latest portion of the dialogue.

   Approaches to improving context management include the use of specialized architectures that compress contextual information more efficiently. For example, models with retrieval-augmented generation (RAG) can effectively access much larger knowledge bases than fit directly in the context window by selectively retrieving relevant information ([45]; see also Gupta, Ranjan, and Singh [46] for a recent survey paper). Similarly, some architectures incorporate separate memory modules or state representations that compress long-term context into compact representations. One promising recent approach is the addition of "memory layers" to the transformer architecture; early results show that memory-augmented models can match the performance of dense models trained with 2–4x more compute on knowledge-intensive tasks [47]. These architectural innovations, if scaled and generalised, may enable adaptation to deployment environments without the computational costs of processing extremely long contexts. Beyond this, we might expect developers to train agentic AI systems that take a greater variety of actions to populate their context windows with relevant information before completing certain tasks; early versions of this include the search functionality now widely integrated into commercial chatbots [48].

3. **Parameter-efficient fine-tuning (PEFT).** Rather than relying solely on context windows to convey local deployment information, general AI systems can have their weights adapted to specific deployment environments through fine-tuning. Traditional fine-tuning is computationally expensive and risks catastrophic forgetting (see section 2.3.2), but newer PEFT methods like LoRA (Low-Rank Adaptation) and its variants allow rapid, efficient adaptation to a specific use case with minimal degradation of base capabilities [49, 50]. These methods can be applied at deployment time with relatively modest computational resources.

   Recent work has demonstrated that PEFT methods can enable effective adaptation to specialized domains, multiple tasks, and even individual user preferences [51, 52, 53]. The "train many, deploy many" paradigm could involve creating specialized adapters for different deployment contexts while maintaining a shared base model. However, many open questions still remain about how well these adaptations generalize beyond their specific training distribution and whether they can handle the full range of contextual variation present in real-world deployments.

Additional evidence that could be used to track progress against this limitation.

- Progress on metrics tracking the long-range memory capability and contextual awareness of mainstream chat assistants and other LLM agents. Existing benchmarks include LoCoMo [54], LongMemEval [55], MemBench [56], Fiction.LiveBench [57], and Michelangelo [58].
- Progress on efficient in-context learning, as tested on simple games like Tic-Tac-Toe, Connect Four, and Gomoku. An existing benchmark is LMAct [59].

[13] Evaluations conducted by AISI researchers for this report have revealed that frontier LLMs likely started playing optimal Tic-Tac-Toe sometime around February 2025, with GPT-5 (full) and Claude 4 models playing optimally, but OpenAI o1, GPT-5 mini/nano, and Claude Sonnet 3.7 all losing a substantial fraction of the time.

2.3.2 Indicator: Continual learning post-deployment

Continual learning is the ability of an AI model to learn from a changing stream of data over time while retaining competence on what it learned before. Current AI systems undergo a self-contained period of training, during which the neural network weights are adjusted and the model gets better at the intended tasks, followed by a period of deployment, during which the AI system is applied to real-world use cases with a largely static capability profile (static apart from whatever can be taught through in-context learning; see below). Unlike humans, transformer-based AI systems arguably do not have a natural mechanism for continual learning throughout their operation.

It is plausible that some form of continual learning is necessary for creating AI systems capable of automating most cognitive labour. An important aspect of human intelligence is the ability to get better at a task over time through a limited amount of practice performing the task. Some experts remarked that it may not be necessary for AI systems to have this ability if we can just train them to be highly capable already prior to deployment. However, it is important to consider that the world is constantly changing, and with it the nature of cognitive labour. Indeed, the automation of work by AI systems is itself a driver of change and could lead to the creation of new tasks for workers that AI systems haven't been trained on. In these conditions, it is hard to imagine how a system with a frozen capability profile could automate all cognitive labour.[14]

For transformer-based AI systems, there are two main categories of methods for implementing continual learning: in-context learning, and periodic retraining. These methods may not be perfectly inter-substitutable and so a comprehensive approach would likely draw on both kinds of methods.

- **Continual learning through in-context learning.** For tasks that an AI model was not trained on, where knowledge of how to do the task is not stored in the model's weights, it is sometimes possible to teach the model how to perform the task by putting a few examples of the target behaviour in the context window. This is called *in-context learning*. Continual learning may be partly solvable through in-context learning, but this is unlikely to be a complete solution insofar as certain skills are intuitively hard to verbalise and/or teach via explicit instructions.

- **Periodic re-training on post-deployment performance data.** The other type of method for implementing continual learning is to collect data from the real-world operation of the model and use this data to periodically re-train it. This could lean on the "train one, deploy many" dynamic whereby data is collected from across all deployed instances of the model and pooled together for (re-)training, and then the resulting update is pushed again to all instances.

Continual learning via periodic re-training is currently bottlenecked by three difficulties: (1) implementation challenges, (2) low sample efficiency, and (3) the threat of "catastrophic forgetting."

1. **Implementation challenges.** These concern when and how often the model is retrained on real-world performance data, as well as issues such as how this data is collected and used. One approach of emerging interest is self-modifying systems [60]: building AI systems that retrain themselves in response to new inputs, without human intervention.

2. **Low sample efficiency.** Current models typically need thousands of training examples to learn intended behaviours. This contrasts with the relatively low rate at which an AI system can be expected to attempt a task and gather performance data during real-world operation. It is possible that by gathering data across all deployed instances, AI systems can make up for this low sample efficiency to some extent, but it seems plausible that technical strategies to promote higher sample efficiency may also be required.

3. **Catastrophic forgetting.** The flip-side of sample-efficient learning is so-called "catastrophic forgetting": if during further training the model drifts too far away from its starting policy, it may then forget other important behaviours that were learned at an earlier stage of training (the canonical reference is French [61]). One expert argued that efficient learning without catastrophic forgetting is a research frontier that might require innovative new approaches, for

[14] Of course, we need not imagine a single deployed model instance automating all cognitive labour for eternity. As the nature of work changes, models can be re-trained on new data. However, this re-training needs to be frequent enough and efficient enough to keep up with the pace of change. Insofar as this can be achieved, frequent re-training may be counted as one method for implementing continual learning.

instance using techniques from mechanistic interpretability.[15] Another expert maintained that this is essentially a solved problem—that is, today's state-of-the-art models don't suffer from catastrophic forgetting any more—but we weren't able to find confirmation of this in the recent literature. Several studies find evidence of catastrophic forgetting when fine-tuning LLMs on downstream tasks [62, 63, 64]. However, there is also evidence that modern LLMs suffer much less from catastrophic forgetting than earlier neural nets, likely due to the intensive pre-training stage [65].

Additional evidence that could be used to track progress against this limitation.

- Demonstrations of new types of training algorithms that achieve markedly higher sample efficiency (e.g., process reward algorithms such as PRIME [19]), especially when combined with low levels of catastrophic forgetting. New algorithms should be tested on models with a reasonably large scale and be benchmarked against widely used baselines (such as PPO/GRPO).
- Scaling of sparse update methods to frontier model sizes (100B+ parameters) with demonstrations of continual learning across diverse domains beyond factual QA, including reasoning, coding, and multi-step task completion [47, 66].
- Progress on self-modifying systems, and evidence of these techniques moving from lab demos to real-world applications. A recent example of a such a technique is Self-Adapting LLMs (SEAL), which teaches LLM-based systems to adapt their own weights based on new inputs [60].

2.4 Originality limitations

2.4.1 Indicator: Original insights

Most interviewed experts remarked that the capacity for original insights appears to be a major failure of current generation AI systems. For example, despite the fact that current frontier models can win gold medals at the International Mathematics Olympiad—a very challenging, highly competitive mathematics competition—to date AI models haven't contributed many novel mathematical concepts or theorems of significant scientific interest.[16] This gap is especially noteworthy if, as one expert remarked, reasoning models have likely produced considerably more tokens of mathematical text over the past half year than all human mathematicians have produced over the same period.[17] Developers have tried to build AI agents that automate the process of scientific discovery (see, e.g., The AI Scientist [69]), but experts were unanimously unimpressed with the results. AI-written research papers, when not beset by many errors, were generally found to be either small variations on existing work, or focussed on random, tangential, or uninteresting hypotheses.

Building AI systems that produce original insights seems to be an unsolved research problem that could mean slower progress to AGI. Most acutely, it could serve as a bottleneck on the trajectory that goes via the automation of AI R&D. According to one popular scenario for AI trajectories, the automation of AI R&D would greatly accelerate progress towards full AGI by effectively increasing the population of AI researchers and the pool of their research ideas (as well as the scientists and engineers needed to test them). But if automated researchers cannot come up with high-quality original ideas to test, then the envisaged acceleration may not happen. More generally, since many jobs require or benefit from original insights, solving this limitation seems necessary to reach AGI.

[15] Mechanistic interpretability is the research programme that attempts to identify the neurons, weights and circuits that implement specific computations in deep neural networks.

[16] Some early signs of progress have emerged over the last few months (May–). For example, Google DeepMind's AlphaEvolve found the first improvement to an algorithm for multiplying two 4x4 matrices in 56 years [67]. In the biomedical domain, FutureHouse's multi-agent system Robin was recently used to generate and validate a novel therapeutic idea for treating dry age-related macular degeneration [68].

[17] This claim strikes us as plausible. In line with community estimates, let's take the population of human mathematicians to be on the order of 100k. If we assume that their average output over the last six months was 200 pages of mathematical text, then this equates to ∼100k tokens per mathematician (at ∼500 tokens/page). So, the entire human maths output for the last six months comes out at around 1e10 tokens. To produce this many tokens in 6 months, reasoning models would need to produce around 55 million tokens of maths per day. OpenAI claims that ChatGPT sees around 2.5 billion prompts per day. If reasoning models use an average of 200–800 tokens (including CoT) in response to the mathematical prompts, then the fraction of prompts which are mathematical queries need be only 0.003–0.011%.

Experts presented a range of ideas concerning the underlying source of the originality limitations. Many of these suggestions would imply that originality might arise alongside progress on other limitations as a result of scaling.

1. **Scaling up existing methods.** At least one expert downplayed originality as a peculiar limitation of current models, attributing the lack of original insights to a simple lack of intelligence. Their argument is that LLM-based AI systems appear more intelligent than they are simply because they have memorised so much more material.[18] If this explanation is correct, then we could perhaps expect originality to come along "by default" as AI systems are made more generally intelligent. Another expert argued that the lack of original insights can be attributed to the short-time horizons over which current systems operate; if true, we may expect originality to arise alongside long-horizon agency (see 2.1.2). A related proposal was that sheer length of thought might substitute for creative inspiration.[19] If this "brute-force Einstein" hypothesis is correct, then we might expect to overcome this limitation through scaling up runtime compute.

2. **New paradigms.** If scaling-up existing methods does not succeed in creating AI systems that generate original insights, then this could point to a more fundamental limitation with the current paradigm. For instance, Francois Chollet has argued that deep learning is inherently limited to a form of interpolation between previously seen data points ("program interpolation"), whereas originality requires more than this (namely, novel "program synthesis") [72].[20] The ARC-AGI benchmarks are designed specifically to test these "programme synthesis" abilities, and indeed the state-of-the-art on ARC-AGI-2 [73] is still much lower than the human baseline. However, overall this was an unpopular view among the interviewed experts, and to our knowledge nothing outside of the deep learning paradigm has had comparable levels of general success.

Additional evidence that could be used to track progress against this limitation.

- Continued improvements on ARC-AGI-2 [73] and other tests of novel "program synthesis" achieved using general-purpose systems (i.e., not using a task-specific solver, and achieved without task-specific fine-tuning).
- Improvements in lab validation rates of AI-originated hypotheses, for instance in functional materials design or pharmaceuticals.
- Improvements in the quality of scientific research conducted end-to-end by AI systems. A clear signal would be increases in the number of fully AI-written research papers that get accepted to top-tier journals and withstand subsequent expert scrutiny. Post-publication durability is also important: another useful signal would be rates of independent replication by human scientists.

3 Conclusion

In this report, we have organised the limitations of current AI systems (as of 09/25) into four broad categories. We find substantial relevant limitations in each category, any one of which might pose a bottleneck for progress towards AI systems capable of automating most cognitive labour. However, we also find substantial evidence of progress across all the limitations. As a result, we conclude that tracking progress against these limitations can provide an informative set of indicators for the AI safety and national security communities to monitor and forecast AI trajectories.

[18] This is somewhat analogous to the difference between a high-performing student who exhaustively studies the past papers for their courses and pattern-matches to the new exams, versus a similarly high-performing student who attempts each problem in the exam without much prior experience of similar problems.

[19] For instance, the mathematician Ramanujan purportedly arrived at his original mathematical insights by filling pages and pages of his notebooks with simple arithmetic. For a description of Ramanujan's distinctive mathematical approach, see any book about Ramanujan, such as Berndt's series on the notebooks [70], or Kanigel's biography [71].

[20] Chollet has shortened his timelines since then, but only due to possible progress on non-LLM systems.

References

[1] Jaime Sevilla et al. *Can AI Scaling Continue Through 2030?* Accessed: 2025-09-24. 2024. URL: https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

[2] Joshua You et al. *Scaling Intelligence: The Exponential Growth of AI's Power Needs*. 2025. URL: https://www.epri.com/research/products/000000003002033669.

[3] *OpenAI Charter*. Accessed: 2025-10-02. 2018. URL: https://openai.com/charter/.

[4] Meredith Ringel Morris et al. "Levels of AGI: Operationalizing Progress on the Path to AGI". In: *CoRR* (2023).

[5] Sayash Kapoor and Arvind Narayanan. *AGI is not a milestone*. Accessed: 2025-10-02. 2025. URL: https://www.normaltech.ai/p/agi-is-not-a-milestone.

[6] Yoshua Bengio et al. *International AI Safety Report*. Tech. rep. DSIT 2025/001. 2025. URL: https://www.gov.uk/government/publications/international-ai-safety-report-2025.

[7] Daniel Eth and Tom Davidson. "Will AI R&D Automation Cause a Software Intelligence Explosion?" In: (2025). Accessed: 2025-09-24. URL: https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.

[8] Daniel Kokotajlo et al. *AI 2027*. Accessed: 2025-09-24. 2025. URL: https://ai-2027.com/.

[9] Jiawei Gu et al. *A Survey on LLM-as-a-Judge*. 2025. arXiv: 2411.15594 [cs.CL]. URL: https://arxiv.org/abs/2411.15594.

[10] *Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad*. Accessed: 2025-10-03. 2025. URL: https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/.

[11] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

[12] Yiping Wang et al. *Reinforcement Learning for Reasoning in Large Language Models with One Training Example*. 2025. arXiv: 2504.20571 [cs.LG]. URL: https://arxiv.org/abs/2504.20571.

[13] Thomas Kwa et al. *Measuring AI Ability to Complete Long Tasks*. 2025. arXiv: 2503.14499 [cs.AI]. URL: https://arxiv.org/abs/2503.14499.

[14] Toby Ord. "Is there a Half-Life for the Success Rates of AI Agents?" In: (2025). Accessed: 2025-09-24. URL: https://www.tobyord.com/writing/half-life.

[15] METR. *Details about METR's evaluation of OpenAI GPT-5*. https://evaluations.metr.org//gpt-5-report/. Aug. 2025.

[16] Hjalmar Wijk et al. *RE-Bench: Evaluating frontier AI RD capabilities of language model agents against human experts*. 2025. arXiv: 2411.15114 [cs.LG].

[17] David Rein et al. *HCAST: Human-Calibrated Autonomy Software Tasks*. 2025. arXiv: 2503.17354 [cs.AI]. URL: https://arxiv.org/abs/2503.17354.

[18] Amrith Setlur et al. *Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning*. 2024. arXiv: 2410.08146 [cs.LG]. URL: https://arxiv.org/abs/2410.08146.

[19] Ganqu Cui et al. *Process Reinforcement through Implicit Rewards*. 2025. arXiv: 2502.01456 [cs.LG]. URL: https://arxiv.org/abs/2502.01456.

[20] *Internal independent review for language model agent alignment*. Accessed: 2025-09-26. 2023. URL: https://www.lesswrong.com/posts/Q7XWGqL4HjjRmhEyG/internal-independent-review-for-language-model-agent.

[21] Noah Shinn et al. *Reflexion: Language Agents with Verbal Reinforcement Learning*. 2023. arXiv: 2303.11366 [cs.AI]. URL: https://arxiv.org/abs/2303.11366.

[22] Aman Madaan et al. *Self-Refine: Iterative Refinement with Self-Feedback*. 2023. arXiv: 2303.17651 [cs.CL]. URL: https://arxiv.org/abs/2303.17651.

[23] Ryo Kamoi et al. "When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs". In: *Transactions of the Association for Computational Linguistics* 12 (2024), 14171440.

[24] Joel Becker et al. *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. 2025. arXiv: 2507.09089 [cs.AI]. URL: https://arxiv.org/abs/2507.09089.

[25] *Project Vend: Can Claude run a small shop? (And why does that matter?)* Accessed: 2025-09-24. 2024. URL: https://www.anthropic.com/research/project-vend-1.

[26] *The AI Village in Numbers*. Accessed: 2025-09-24. 2025. URL: https://theaidigest.org/village/blog/village-in-numbers.

[27] Adam Tauman Kalai et al. *Why Language Models Hallucinate*. 2025. arXiv: 2509.04664 [cs.CL]. URL: https://arxiv.org/abs/2509.04664.

[28] Karl Cobbe et al. *Training Verifiers to Solve Math Word Problems*. 2021. arXiv: 2110.14168 [cs.LG]. URL: https://arxiv.org/abs/2110.14168.

[29] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=d7KBjmI3GmQ.

[30] David Rein et al. *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. 2023. arXiv: 2311.12022 [cs.AI]. URL: https://arxiv.org/abs/2311.12022.

[31] Dongxu Zhang et al. *ASCoT: An Adaptive Self-Correction Chain-of-Thought Method for Late-Stage Fragility in LLMs*. 2025. arXiv: 2508.05282 [cs.CL]. URL: https://www.arxiv.org/abs/2508.05282.

[32] Trieu H. Trinh et al. "Solving olympiad geometry without human demonstrations". In: *Nature* 625 (2024), 476482.

[33] *HHEM 2.1: A Better Hallucination Detection Model and a New Leaderboard*. Accessed: 2025-09-29. 2024. URL: https://www.vectara.com/blog/hhem-2-1-a-better-hallucination-detection-model.

[34] Junyi Li et al. "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. DOI: 10.18653/v1/2023.emnlp-main.397. URL: https://aclanthology.org/2023.emnlp-main.397/.

[35] Yejin Bang et al. *HalluLens: LLM Hallucination Benchmark*. 2025. arXiv: 2504.17550 [cs.CL]. URL: https://arxiv.org/abs/2504.17550.

[36] *Expanding on what we missed with sycophancy*. Accessed: 2025-09-29. 2025. URL: https://openai.com/index/expanding-on-sycophancy/.

[37] METR. *Update on ARC's recent eval efforts*. https://metr.org/blog/2023-03-18-update-on-recent-evals/. Mar. 2023.

[38] Saurav Kadavath et al. *Language Models (Mostly) Know What They Know*. 2022. arXiv: 2207.05221 [cs.CL]. URL: https://arxiv.org/abs/2207.05221.

[39] Jan Betley et al. *Tell me about yourself: LLMs are aware of their learned behaviors*. 2025. arXiv: 2501.11120 [cs.CL]. URL: https://arxiv.org/abs/2501.11120.

[40] Yoo Yeon Sung et al. *GRACE: A Granular Benchmark for Evaluating Model Calibration against Human Calibration*. 2025. arXiv: 2502.19684 [cs.CL]. URL: https://arxiv.org/abs/2502.19684v1.

[41] Ezra Karger et al. *ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities*. 2025. arXiv: 2409.19839 [cs.LG]. URL: https://arxiv.org/abs/2409.19839.

[42] Greg Burnham and Tom Adamczewski. *LLMs now accept longer inputs, and the best models can use them more effectively*. Accessed: 2025-09-26. 2025. URL: https://epoch.ai/data-insights/context-windows.

[43] Nelson F. Liu et al. *Lost in the Middle: How Language Models Use Long Contexts*. 2023. arXiv: 2307.03172 [cs.CL]. URL: https://arxiv.org/abs/2307.03172.

[44] Cheng-Ping Hsieh et al. *RULER: What's the Real Context Size of Your Long-Context Language Models?* 2024. arXiv: 2404.06654 [cs.CL]. URL: https://arxiv.org/abs/2404.06654.

[45] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[46] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions*. 2024. arXiv: 2410.12837 [cs.CL]. URL: https://arxiv.org/abs/2410.12837.

[47] Vincent-Pierre Berges et al. *Memory Layers at Scale*. 2024. arXiv: 2412.09764 [cs.CL]. URL: https://arxiv.org/abs/2412.09764.

[48] *Introducing ChatGPT search*. Accessed: 2025-10-22. 2024. URL: https://openai.com/index/introducing-chatgpt-search/.

[49] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[50] Ning Ding et al. "Parameter-efficient fine-tuning of large-scale pre-trained language models". In: *Nature machine intelligence* 5.3 (2023), pp. 220–235.

[51] Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. "Parameter-Efficient Legal Domain Adaptation". In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. Ed. by Nikolaos Aletras et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 119–129. DOI: 10.18653/v1/2022.nllp-1.10. URL: https://aclanthology.org/2022.nllp-1.10/.

[52] Yaming Yang et al. "MTL-LoRA: Low-Rank Adaptation for Multi-Task Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 20. 2025, pp. 22010–22018.

[53] Zhaoxuan Tan et al. *Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning*. 2025. arXiv: 2402.04401 [cs.CL]. URL: https://arxiv.org/abs/2402.04401.

[54] Adyasha Maharana et al. "Evaluating Very Long-Term Conversational Memory of LLM Agents". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13851–13870. DOI: 10.18653/v1/2024.acl-long.747. URL: https://aclanthology.org/2024.acl-long.747/.

[55] Di Wu et al. *LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory*. 2025. arXiv: 2410.10813 [cs.CL]. URL: https://arxiv.org/abs/2410.10813.

[56] Haoran Tan et al. "MemBench: Towards More Comprehensive Evaluation on the Memory of LLM-based Agents". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 19336–19352. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.989. URL: https://aclanthology.org/2025.findings-acl.989/.

[57] *Introducing Fiction.LiveBench: The First Real-World Long Context Benchmark for Writers*. Accessed: 2025-09-26. 2025. URL: https://fiction.live/stories/Fiction-liveBench-Feb-21-2025/oQdzQvKHw8JyXbN87.

[58] Kiran Vodrahalli et al. *Michelangelo: Long Context Evaluations Beyond Haystacks via Latent Structure Queries*. 2024. arXiv: 2409.12640 [cs.CL]. URL: https://arxiv.org/abs/2409.12640.

[59] Anian Ruoss et al. *LMAct: A Benchmark for In-Context Imitation Learning with Long Multimodal Demonstrations*. 2025. arXiv: 2412.01441 [cs.AI]. URL: https://arxiv.org/abs/2412.01441.

[60] Adam Zweiger et al. *Self-Adapting Language Models*. 2025. arXiv: 2506.10943 [cs.LG]. URL: https://arxiv.org/abs/2506.10943.

[61] Robert M. French. "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Sciences* 3.4 (1999), pp. 128–135.

[62] Yun Luo et al. "An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning". In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 3776–3786. DOI: 10.1109/TASLPRO.2025.3606231.

[63] Damjan Kalajdzievski. *Scaling Laws for Forgetting When Fine-Tuning Large Language Models*. 2024. arXiv: 2401.05605 [cs.CL]. URL: https://arxiv.org/abs/2401.05605.

[64] Haizhou Shi et al. *Continual Learning of Large Language Models: A Comprehensive Survey*. 2024. arXiv: 2404.16789 [cs.LG]. URL: https://arxiv.org/abs/2404.16789.

[65] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. "Fine-tuned Language Models are Continual Learners". In: *Conference on Empirical Methods in Natural Language Processing*. 2022. URL: https://api.semanticscholar.org/CorpusID:252815378.

[66] Jessy Lin et al. *Continual Learning via Sparse Memory Finetuning*. 2025. arXiv: 2510.15103 [cs.CL]. URL: https://arxiv.org/abs/2510.15103.

[67] Alexander Novikov et al. *AlphaEvolve: A coding agent for scientific and algorithmic discovery*. 2025. arXiv: 2506.13131 [cs.AI]. URL: https://arxiv.org/abs/2506.13131.

[68] Ali Essam Ghareeb et al. *Robin: A multi-agent system for automating scientific discovery*. 2025. arXiv: 2505.13400 [cs.AI]. URL: https://arxiv.org/abs/2505.13400.

[69] Chris Lu et al. *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. 2024. arXiv: 2408.06292 [cs.AI]. URL: https://arxiv.org/abs/2408.06292.

[70] Bruce C. Berndt. *Ramanujans Notebooks*. Springer New York, 2012.

[71] Robert Kanigel. *The man who knew infinity*. Simon and Schuster, 2013.

[72] *Francois Chollet Why the biggest AI models can't solve simple puzzles*. Accessed: 2025-09-26. 2024. URL: https://www.dwarkesh.com/p/francois-chollet.

[73] Francois Chollet et al. *ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems*. 2025. arXiv: 2505.11831 [cs.AI]. URL: https://arxiv.org/abs/2505.11831.