Continual Learning Requires Evaluating Trajectories

Lorenzo Pacchiardi[1,*] **Patricia Paskov**[2] **Seán Ó hÉigeartaigh**[1]

Fernando Martínez-Plumed[3] **Katherine M. Collins**[4] **Fazl Barez**[2] **Jonathan Prunty**[1]

Matteo Gabriel Mecattaf[1] **Zafeirios Fountas** **Risto Uuk**[5]

Sanmi Koyejo[6,†] **Cozmin Ududec**[†]

José Hernández-Orallo[1,3,†]

[1]University of Cambridge [2]University of Oxford [3]Universitat Politècnica de València
[4]MIT [5]Future of Life Institute [6]Stanford University

Abstract

AI systems increasingly incorporate continual learning mechanisms allowing their behaviour to adapt after deployment, from in-context learning and memory features already in wide use to post-deployment weight modification under research. We argue that, by treating AI systems as frozen artefacts whose performance and safety are assessed at release, current evaluation practices structurally ignore the behavioural trajectory of a system that continues to learn from experience. Our position is that evaluation of continual learning systems should be centred on behavioural trajectories, with the complementary goals of characterising the landscape of possible behaviours and forecasting how behaviour will evolve from a given set of experiences. This can be operationalised through trajectory elicitation sandboxes and monitors that forecast behavioural evolution, but may face fundamental obstacles analogous to those seen in dynamical systems. These are best addressed by applying trajectory-centred evaluation to today's continual learning systems and relying on the resulting evidence to design systems amenable to it, yielding a virtuous cycle in which systems and their evaluations co-evolve.

1 Introduction

The standard depiction of frontier AI models describes them as trained before deployment, released with frozen parameters, and as carrying no information across interactions, so that multiple instances produce the same probability distribution of outputs for a given input across users and over time [16].

This temporal and user invariance is rapidly eroding. Models such as Large Language Models (LLMs) [105, 72] alter their response based on previous turns of a persistent session [17] and increasingly benefit from massive context windows that allow them to incorporate a larger number of turns [96]. Moreover, production LLMs [105, 72] are wrapped in "systems" that equip them with "memory" persisting *across sessions* [78, 5] and can write and retrieve external stores [4, 60]. Active research aims to allow AI systems to modify their parameters (e.g., neural network weights) in response to deployment-time experiences [31, 39, 12, 49, 35]. These mechanisms are instantiations

*Corresponding author: lp666@cam.ac.uk.
†Equal senior authorship.

Preprint.

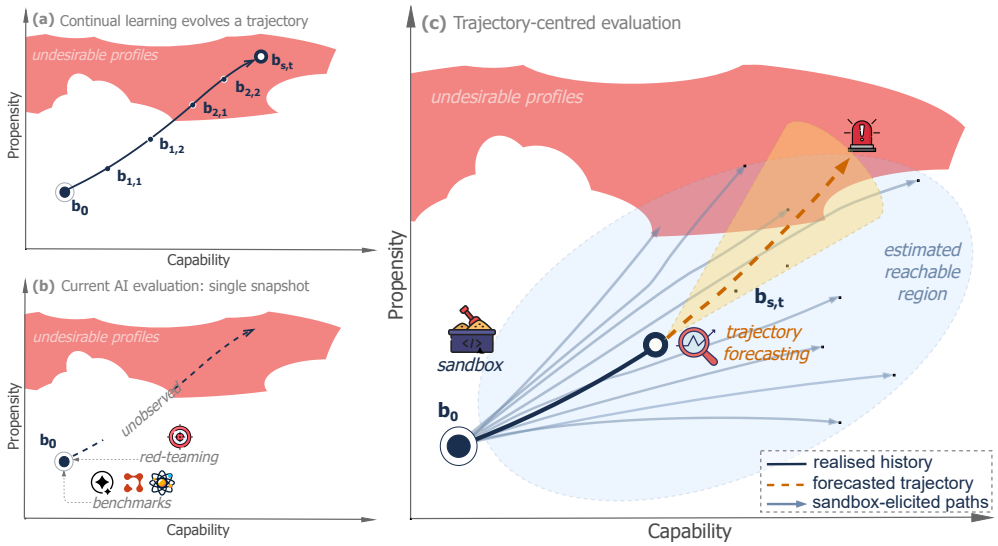

Figure 1: A CL AI system can be described by a behavioural profile encompassing capabilities and propensities. (a): the profile b_0 describing the system at release time evolves over turns t (within context) and sessions s (across context) generating a trajectory $b_{s,t}$ (shown in a simplified bi-dimensional space). (b): current AI evaluation relies on benchmarking and red-teaming the released system and is therefore inadequate to observe realised trajectories for systems that learn after deployment. (c): our vision of evaluation for continual learning combines pre-deployment sandbox elicitation of trajectories, to characterise the landscape of behavioural space the system can reach, and post-deployment forecasting of future trajectories, to predict development of undesirable profiles.

of "continual learning" (CL, also referred to as lifelong learning [24], or incremental learning [97]) as they share a structural property: deployed system instances learn over deployment time and (potentially) specifically to a user, along trajectories shaped by experience.

This feature undermines current evaluation practices, which rely on temporal and user invariance: a result obtained through benchmarks [40, 22, 101] or red-teaming [36] on the released checkpoint is taken to characterise the deployed model facing all users, with no information automatically retained across interactions. In that paradigm, any change to the system requires a deliberate act by the developer or a user (e.g., a new training run or "fine-tuning" [29, 45, 44]).

In this paper, we argue that the emergence of systems that automatically learn over time and break temporal and user invariance requires rethinking evaluation practice. Thus, **our position is that evaluation of continual learning systems should be centred on behavioural trajectories, with the complementary goals of characterising the landscape of possible behaviours and forecasting how behaviour will evolve from a given set of experiences** (Fig. 1). To this end, in Sec. 2, we discuss the levels of CL and how CL describes trajectories. Then, in Sec. 3, we discuss undesirable behavioural changes and why current evaluation practices are unequipped to address these. In Sec. 4, we present our vision for the goals of evaluation for CL and how that can be operationalised with trajectory elicitation sandboxes and predictive monitors. Sec. 5 then highlights potential issues that may hinder our vision and Sec. 6 advocates for starting to build trajectory-centred evaluation on today's systems to accumulate evidence and expertise and for developers to design CL systems amenable to evaluation. Sec. 7 considers alternative views and Sec. 8 concludes.

2 Setting the ground: definitions and concepts

An AI model is "a computational construct that makes inferences from inputs to produce outputs", while an AI system encompasses "one or more AI models, an interface for receiving inputs from and delivering outputs to an environment, and the configuration connecting these" [95]. An AI "agent" is an AI system enabling interaction with a (virtual) environment. We use "AI systems" throughout and consider individual AI models as embedded within a minimal wrapping system. An AI system is

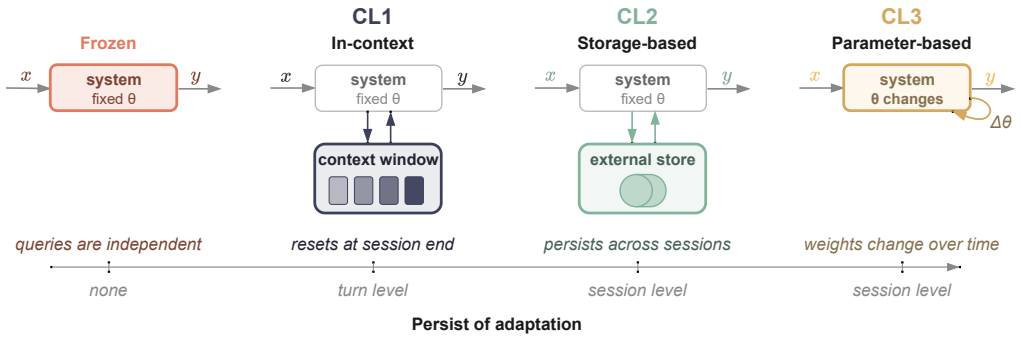

Figure 2: AI systems map input x to output y. Frozen systems embed no adaptation across turns or sessions, so that the output distribution is invariant over time and users. CL1 allows information across turns to accumulate within a session without preservation across sessions. CL2 adds an external store that persists across turns and sessions while weights stay fixed, and CL3 allows the system parameters themselves to change. The different levels can be combined in different ways.

defined by a set of parameters—such as neural network weights or how AI models within the system are connected. We use "environment" to refer to everything external to the AI system: users, other AI systems, and beyond. We consider general-purpose AI models which encode inputs into a "context window" and produce outputs stochastically. Often, a model's output to a user query can be appended to the original query within the same context window, enabling multiple "turns" of interactions. Interactions happening in separate context windows are referred to as different "sessions".

2.1 Levels of continual learning

We distinguish three levels of mechanisms giving rise to CL[3] based on the object through which adaptation emerges (see Fig. 2 for a representation):

In-Context Continual Learning (CL1): The system progressively accumulates information over turns within a single session and this affects the way it responds, but does not persist beyond the session. This is widely used in LLM apps (such as ChatGPT) or agent frameworks [100].

Storage-based Continual Learning (CL2): The system has an external store where it can write information and access it in later turns or sessions. Examples include memory in chatbot apps (as in multisession ChatGPT, [23]), "agent skills" [4], Retrieval Augmented Generation (RAG, [60]), agentic systems with file editors [100], or approaches that iteratively edit their own context [104].

Parameter-based Continual Learning (CL3): System's parameters are modified post-deployment in response to experiences, persisting across turns and sessions. CL3 includes methods modifying the underlying models' weights [31, 39, 12, 49, 35, 48] and the agent's architecture itself [102].

In contrast, we refer to a system as "frozen" if it incorporates none of these mechanisms, so that its behaviour does not adapt across turns or sessions.

Some approaches may be situated at the intersection of different levels: in several RAG approaches [60], the retrieval index is updated over time (part of CL3). More generally, systems can combine multiple levels of CL. While CL1 is naturally applied at the level of individual users, CL2 and CL3 can be performed independently over different users or centralised by updating a single model.

Our taxonomy is agnostic to the algorithm by which learning occurs (we refer to Khetarpal et al. for this [55]) and is suitable for CL in "narrow" AI systems that are sequentially trained on one new task at a time [24, 99], as is traditional in reinforcement learning (RL) paradigms [55]. However, our main focus is CL for general-purpose AI systems, which excludes one-off mechanisms such as post-deployment user-initiated fine-tuning or new version releases from AI providers. These do not possess the continuous and autonomous adaptation to experience that motivates our work.

[3]Other AI paradigms (e.g., without context windows) may be better characterised by different levels of CL.

2.2 Statefulness in continual learning

All CL levels include information from past experiences into present behaviour through an evolving internal state, which we refer to as *statefulness*. Formally, we describe a CL AI system after turn t within session s by an internal state $z_{s,t}$, which may include previous inputs and outputs, memories, model weights, system parameters, and other features. The state evolves according to an update mechanism based on received input $x_{s,t}$, produced output $y_{s,t}$, and the environment feedback $r_{s,t}$, that depends on the level(s) of CL in operation. At any time, the system's output to input x is drawn from the conditional distribution $p(\cdot \mid z_{s,t}, x)$ over the space of outputs the system is able to generate. Through dependence on $z_{s,t}$, the same query, posed at different sessions and turns, may produce different output distribution. For a frozen system, $z_{s,t} = z_{0,0}$ for all t and s, so that the output distribution depends only on the input x. This formalism is a stripped-down Partially Observable Markov Decision Process [53]. We do not model the environment's transition dynamics in detail: the system's update rule/mechanism, not the environment's, is our object of interest. Adjacent formalisms include online learning [41], contextual bandits [64], and lifelong-learning Markov decision processes [24, 55].

2.3 State and behavioural profiles

The state $z_{s,t}$ is inconvenient as an object of evaluation: it is high-dimensional, incommensurable across systems, and most of its dimensions do not directly correspond to features of behaviour. A similar problem is faced in psychology: it is unwieldy to describe a human's behaviour by directly referring to their past experience and genotype. Instead, psychology relies on *traits*: relatively stable individual differences that persist over time and apply across tasks [73, 85, 69].

Analogously, we work in a lower-dimensional, system-agnostic projection of $z_{s,t}$ that we call the system's *behavioural profile* $b_{s,t}$ which allows us to describe the behaviour of a system over a wide range of inputs.In $b_{s,t}$ we include *capabilities*, which characterise what the system can potentially do, encompassing both abstract cognitive proficiencies (e.g., planning depth) and domain-specific proficiencies (e.g., assisting chemical synthesis), and *propensities*, which characterise what the system tends to do under different instigation conditions [86], such as refusal patterns, honesty, interactional styles, and decision-making dispositions.

While different choices of "coordinate systems" for capabilities and propensities are possible [107, 42, 19], no behavioural profile can fully predict a model's behaviour. This is due to the residual epistemic uncertainty, arising from our inability to fully define a system's underlying constructs, and the irreducible aleatoric uncertainty arising from the model's stochastic nature. Two additional elements add further complexity: *factual knowledge* $K_{s,t}$, i.e., pieces of information part of the state $z_{s,t}$ identified by reference (e.g., a memorised API key or a customer's billing record), and *affordances* $A_{s,t}$, i.e., tools, permissions, integrations, and other deployment-context features that enable actions. Factual knowledge and affordances condition system behaviour in specific scenarios and cannot be aggregated into low-dimensional traits affecting behaviour more broadly.

2.4 Trajectories of continual learning AI systems

Over turns and sessions, the internal state and behavioural profile of a CL system describe *trajectories* depending on the environment and the system's update mechanism. We visualise this for the profile in a simplified space of one capability and one propensity in Fig. 1(a). Different instances of a released system encounter different users, tasks, and tool outputs, thus leading to diverging trajectories.

Crucially, the inputs presented by the environment can be affected by the previous output of the system. This can consist of the environment — including human users — reacting to the system's output in producing the next input, or of the system itself actively choosing the next input as part of its output (e.g., which tool to invoke). These passive and active *non-stationary* scenarios (Sec 4.8 in [55]) contrast with cases in which the sequence of inputs is produced by a *stationary* distribution. While both generate trajectories, the non-stationary case is more complex due to the presence of a feedback loop, and is more likely to give rise to the obstacles we discuss in Sec. 5.

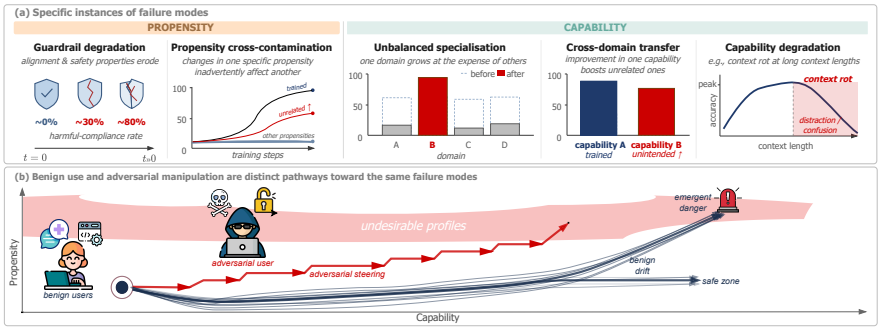

Figure 3: (a): Example failure modes of CL: propensity changes include guardrail degradation and cross-contamination, where altering one propensity inadvertently affects another one. Capability changes include specialisation in one capability at the expense of others, cross-domain transfer of capability gains, or degradation of capability over turns or sessions. (b): Undesirable capability and propensity changes can occur through unintended divergence from benign use or adversarial manipulation of experiences, with both potentially leading to undesirable behavioural profiles.

3 Current evaluation practice is insufficient for continual learning systems

3.1 Continual learning may cause undesirable capability and propensity changes

As a CL system learns from experience, its capabilities and propensities may change in undesirable ways that lead the system to produce harmful outputs, reduce its usefulness (if it becomes unable or unwilling to perform tasks), or increase the jaggedness of its behavioural profile and thus make the system hard to use reliably [74]. Examples of failure modes are shown in panel (a) of Fig. 3: for propensities, these consist of *guardrail degradation*, where alignment and safety properties erode over time, and *propensity cross-contamination*, where changes in one specific propensity inadvertently affect another, seemingly unrelated, one. For capabilities, they include *unbalanced specialisation*, where the system becomes proficient in a specific domain at the expense of others; *cross-domain transfer*, where improvement in one capability boosts seemingly unrelated capabilities; and *capability degradation*, where capabilities erode after a large number of turns or sessions.

CL1 already exhibits several of these failure modes in current LLMs—capability degradation [59] and guardrail degradation enabling jailbreaks [87, 37]—and their emergence in CL2 and CL3 is similarly plausible: empirical analogues appear even in the controlled setting of one-off user-directed training, and CL's gradual cumulative nature only expands the surface area over which failures can emerge. Specifically, propensity cross-contamination occurred when rewarding a "Nerdy" conversational personality amplified a model's tendency to mention goblins [79] and when fine-tuning on specific misaligned behaviours produced "emergent misalignment" in broader contexts [13]. Unbalanced specialisation has been extensively documented as "catastrophic forgetting" in domain-specific fine-tuning [65, 93]. Cross-domain transfer is exemplified by the finding that including code in pre-training data improves natural-language reasoning [7]. Finally, capability degradation happened with fine-tuning that encourages hallucinations and reduces contextual reasoning [54, 38].

Undesirable changes in capabilities and propensities can materialise through (panel b of Fig. 3):

Unintended benign-use divergence: for instance, a medical system learning from empathetic dialogues losing its ability to reason about patient privacy; a coding system assisting a security researcher developing fluency in exploit techniques as a side effect of improving at the legitimate tasks.

Adversarial manipulation: deliberate steering of the learning process (by an external actor or the CL system itself) toward undesirable behaviours; for instance, an attacker could craft a sequence of seemingly harmless chemistry tutoring requests that incrementally shift the model's propensities around safety refusals or steer its specialisation toward exploit-relevant capabilities.

On top of unintended or adversarial undesirable changes in propensities and capabilities, any snapshot of a CL system shares the concerns posed by frozen AI systems. However, AI evaluation practice targets the latter—although imperfectly—while being inadequate to address issues intrinsic to CL.

3.2 Pre-deployment evaluation of AI systems assumes they do not change after deployment

Current practice [101, 22, 40, 18] conducts pre-deployment evaluations on the snapshot to be released and various fine-tuned and pre-mitigation (e.g., without harmlessness training) snapshots; instruments span single-turn benchmarks [40, 22], multi-turn conversations [59, 37], agent evaluations [101], and red-teaming [36]. The results are taken to be informative of the deployed system's behaviour under most circumstances it will encounter. Therefore, **the current evaluation practice assumes that AI systems do not change after deployment** or at most undergo user-initiated fine-tuning emulated by pre-mitigation studies. **This makes it structurally inadequate to target the undesirable propensity and capability changes intrinsic to CL AI systems**[4]**.** On top of these fundamental concerns with CL, AI evaluation has well documented methodological and validity issues [10, 71, 14, 52, 84, 77, 58, 91], which affect evaluations' informativeness for real-world behaviour, and they can compounded by CL systems: even a small change in behaviour can make an already fragile estimate useless.

Pre-deployment evaluation is complemented in current practice by deployment-time monitoring and control, such as input/output [67, 50, 83] and chain-of-thought (CoT) monitors [8, 57]. These monitors are calibrated on pre-deployment evaluations (e.g., which inputs to filter, which CoT patterns to flag) and therefore inherit the core assumption that the system is a fixed artefact. This introduces correlated failure modes: a CoT monitor calibrated on the released checkpoint may over-fire on inputs the system has since learned to handle correctly, or miss novel failure modes that emerge as its reasoning evolves.

Existing benchmarks for measuring CL failure cases are limited to specific domains for CL1 [46, 20] and CL2 [106] and remain isolated [3, 32]. Similarly, RL-based CL environments [55] are mostly domain-specific or report metrics about generalisation, transfer, and overall performance rather than the trajectory of the behavioural profile. At the same time, RL-flavoured control CL approaches such as shielding [1] and constrained MDPs [2] require a formalism that does not transfer cleanly to general-purpose models undergoing CL. Thus, CL systems require evaluation to consider the behavioural trajectories shaped by the experiences each deployed instance encounter and develop integrated mechanisms for both their generation and prediction, as we discuss next.

4 Evaluation for Continual Learning Systems

4.1 The goals of evaluation for continual learning AI systems

Ideally, a deployed CL system should be tested after each adaptation step to ensure its behaviour remains desirable [3, 32]. However, this is infeasible with systems undergoing frequent and user-specific adaptation. Thus, we argue that evaluation should adopt the two complementary goals of understanding the possible behaviours a CL system can develop and monitor the system's trajectories:

Landscape characterisation relies on realised trajectories to map the behavioural profiles a CL system can reach and the probability it does so under various conditions. Specific questions include:

- Over the distribution of typical input sequences, what is the probability that the system develops undesirable (e.g., with low capabilities or unsafe propensities) behavioural profiles?
- Once the system has reached a benign behavioural profile, how easy is it for adversaries to cause the system to move to an undesired profile?
- Can a system's capability and propensities grow indefinitely, and if not, what are the extreme values they could develop?

Trajectory forecasting aims to predict how the trajectory of a CL system will progress. Considering a system in state $z_{s,t}$ receiving a given input (or a sequence thereof), specific questions include:

- Can we produce calibrated forecasts of the future behavioural profile $b_{s,t+h}$ at horizon h?
- What is the probability that the resulting behavioural profile will be undesirable, or more easily cause the development of such profiles in the future?

These two goals address unintended benign-use divergence and adversarial manipulation (Sec. 3.1) in complementary ways. *For benign divergence*, landscape characterisation is primary in understanding the probability of ordinary use leading to different regions of behavioural space. Trajectory forecasting is complementary, flagging whether a deployed instance's experiences steer it toward an undesirable profile early enough to intervene, before the trajectory becomes harder to forecast. *For adversarial*

[4] See [9] for an accessible discussion of how this break manifests in concrete deployment scenarios.

manipulation, the priority reverses: trajectory forecasting is primary in detecting whether the current sequence of inputs originates from adversarial attempts to reach undesirable profiles. Landscape characterisation complements this by pre-emptively characterising dangerous states and how close they are to typical deployment states, which reveals how much adversarial effort is needed to reach them.

4.2 Operationalising continual learning evaluation

The two goals can be operationalised through infrastructure to elicit controlled trajectories of CL systems and developing monitors to predict evolution of behavioural profiles.

Trajectory elicitation sandboxes AI evaluators should build *trajectory elicitation sandboxes* where a CL system is subjected to prolonged controlled interactions across multiple sessions in parallel. In the sandbox, it should be possible to generate the input sequences in multiple ways: *first*, prescribed by evaluators, which allows them to systematically test multiple conditions and conduct adversarial stress-testing, as in procedural test generation for frozen systems [51, 82, 103] and red-teaming for CL1 systems [36]. *Second*, by emulating environment and users [63, 62, 61] that realistically react to the system's outputs, so that sequences realistically represent real-world interactions; this corresponds to the passive non-stationary scenario of Sec. 2.4. *Third*, by accommodating various levels of self-directedness a system can display [3, 55] and allow it to actively choose the next input based on previous interactions; this corresponds to the active non-stationary scenario of Sec. 2.4. In the latter case, the sandbox should be aware of the risk of strategic behaviour [58] of the tested system (e.g., altering its exploration once aware of being in an evaluation environment [77]).

The sandbox should also allow for various ways to generate feedback to the system's output: rule-based, human (emulated) feedback, or self-generated reflection of the system itself. After each round of feedback, the evaluated system should be able to evolve with the CL levels it employs; where a system employs multiple CL approaches, the combination should be accounted for. At regular intervals along each elicited trajectory, the sandbox should suspend the system's learning ability and subject it to capability and propensity evaluations (e.g., benchmarks or red-teaming approaches) to estimate the current behavioural profile and chart its evolution over interactions.

Predictive monitors The trajectories elicited in sandbox can be leveraged to train *predictive monitors* that consider a system's current state $z_{s,t}$ and one or more upcoming inputs from the environment (possibly spread across multiple sessions), and forecasts the behavioural profile the system will develop or specific features of it (e.g., whether a given capability will exceed a threshold). Forecasts may target the immediate next step or a longer horizon. The latter constitutes a second-order prediction problem in which, for non-stationary scenarios (Sec. 2.4), the monitor must implicitly consider how the system outputs or its active exploration shape its future inputs. Crucially, predictive monitors should provide calibrated uncertainty quantifications and flag when a trajectory is becoming less predictable or likely originated from adversarial manipulation (similarly to anomaly detection [21]).

A precise mechanistic understanding of the future internal state is not necessary; instead, predictive monitoring can be framed as a standard supervised learning problem: each elicited trajectory yields a sequence of tuples (state, future interactions, measured $\hat{b}_{s,t}$) at various prediction horizons, which can be used to train and validate the monitor within the sandbox before deployment. An example of this for CL1 was developed by Bigelow et al. [15], who employed a Bayesian model to predict an LLM's behaviour as in-context evidence accumulates.

5 Obstacles to the goals of evaluation for continual learning AI systems

Potential fundamental obstacles exist to achieve the goals of evaluation for CL systems.

Sensitivity to states and inputs. Predictive monitors rest on a smoothness assumption: small differences in the system's state or the inputs it receives produce correspondingly small differences in the resulting behavioural profile. Chaos theory [94] shows that many deterministic dynamical systems violate this assumption through *sensitive dependence on initial conditions*: two states differing by an arbitrarily small perturbation diverge at an exponential rate quantified by the system's "Lyapunov exponents". Beyond a characteristic time horizon—depending on the precision with which the initial state is known—prediction becomes intractable even given perfect knowledge of the update

rule. Weather forecasting illustrates this regime: the governing equations are well understood, yet long-range prediction fails because of exponential sensitivity to unmeasurable details.

An analogous regime is plausible for CL systems: token-mixing operations in CL1 and memory-management operations in CL2 are high-dimensional non-linear functions, and the loss landscapes of CL3 parameter updates contain saddle points and narrow valleys associated with chaotic dynamics [28]. Stochasticity of transitions does not help: the theory of random dynamical systems [6] establishes that nonlinear stochastic systems admit Lyapunov exponents analogous to their deterministic counterparts, so small perturbations to the state or input sequence may be exponentially amplified along a trajectory regardless of whether transitions are deterministic. [5] Ecosystem coupling can in principle stabilise the dynamics by steering the trajectory onto otherwise unstable periodic orbits [80]. On the other hand, coupling may generate further instability, since coupled learner–environment systems are known to exhibit quasiperiodicity, limit cycles, intermittency, and deterministic chaos even in very simple reinforcement-learning settings [88, 89]. The implication is that, where chaotic dynamics arise, sandbox trajectories are insufficient to train predictive monitors that perform well on deployment trajectories. The epistemic burden is asymmetric: chaos can be *demonstrated* from a single divergent trajectory pair, but its *absence* requires a global guarantee that no region of the state space exhibits sensitive dependence, and absence of chaos in the regions sampled by a sandbox does not imply absence in the regions encountered during deployment.

Multiplicity of attractors. Landscape characterisation relies on the trajectories elicited in pre-deployment sandboxes being informative of the full range of behavioural profiles a system may develop. This fails when the system, viewed as a dynamical system, has multiple *attractors*—regions of state space toward which trajectories starting in their respective *basins* converge, while trajectories starting elsewhere never enter. Hopfield networks [43] illustrate this: a recurrent neural network whose weights are chosen so that a collection of patterns are fixed points of its dynamics, with each stored pattern surrounded by its own basin. A single trajectory, or a collection of trajectories initialised within one basin, reveals nothing about the others.

An analogous regime is plausible for CL AI systems: the loss landscapes of CL3 parameter updates contain many local minima [25] that can act as attractors; in CL2, previously stored content biases retrieval and can self-reinforce; in CL1, long contexts can lock behavioural regimes in [92]. Here too, stochasticity and coupling cut both ways: stochastic transitions may enable a system to escape shallow attractors, analogously to thermal fluctuations in simulated annealing [56], but a reactive ecosystem can equally deepen them, as when user feedback selectively reinforces outputs characteristic of the attractor and raises the barrier to escape. The epistemic burden is similarly asymmetric: demonstrating an additional attractor requires a single trajectory that converges to it, whereas establishing that the attractors identified in the sandbox are the only ones requires a global statement about the entire state space, which is hard due to the potentially fractal nature of their basins [70]. A sandbox in which every elicited trajectory converges to a benign attractor is compatible with the system being genuinely well-behaved globally and the system admitting rare input sequences that drive it into undesirable basins.

The adversarial case worsens both obstacles. An attacker is unconstrained to benign usage patterns and can use input sequences that exploit either pathology: driving the system into chaotic regions where predictive monitors lose grip, or crossing into poorly characterised basins. In either case, the undesirable behavioural profile may persist long after the triggering input sequence has ended.

Chaoticity and multiplicity of attractors are independent: a system can display only the former (e.g., Lorenz's system [94]), only the latter (Hopfield network) or both (Chua's circuit [26]).

6 Addressing the core obstacles

The obstacles identified in Sec. 5 are avoidable: their prevalence and severity for a given system are empirical questions. We recommend two complementary lines of action to address them: evaluators can gather evidence on where and how they arise by applying the tools of Sec. 4.2 to existing CL systems and developers can, aware of these obstacles, design systems that preserve evaluability.

Start trajectory-centred evaluation on today's systems. A few benchmarks test CL1 and CL2 systems in narrow scenarios [46, 20, 106] and some environments provide feedback to LLM agents on

[5] Stochasticity can also cause a system to converge to a stationary measure [68], which guarantees that long-run statistical properties of the system are independent of initial conditions, but says nothing about finite-horizon behaviour

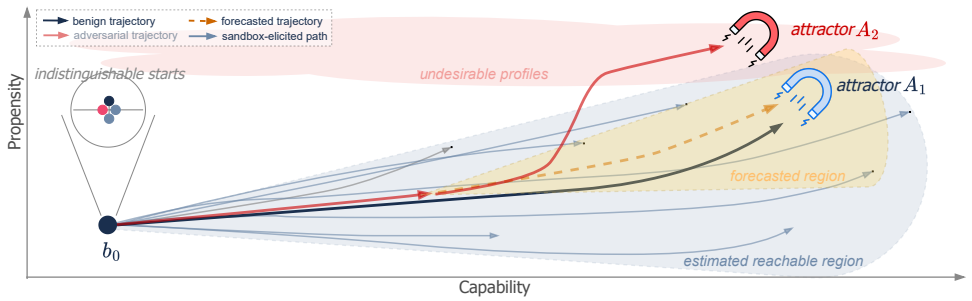

Figure 4: Fundamental obstacles may hinder trajectory-centred evaluation: chaotic sensitivity to states and initial conditions may prevent predictive monitors to generalise from elicited trajectories, while multiplicity of attractors may cause elicited trajectories to represent only a subset of the possible behavioural profiles and deployed systems may converge to attractors that are hard to identify.

research tasks [76, 75]. Evaluators should extend these into trajectory elicitation sandboxes covering a comprehensive set of tasks, satisfying the design dimensions of Sec. 4.2, and extensible to future CL mechanisms. The resulting data should be used to build predictive monitors for current systems and investigate their performance. Together, these initial experiments would provide empirical signal on the obstacles raised in Sec. 5, thus yielding the evidence on which the next recommendation depends.

Design CL mechanisms amenable to evaluation. Model developers should identify, through theoretical arguments and evidence from initial experiments (recommendation above), CL mechanisms that prevent the obstacles in Sec. 5 from arising. Such systems are more amenable to evaluation and should therefore be preferred. Four design directions seem promising. *(i) Contractive update rules* [66] yield predictable long-horizon trajectories. *(ii) Intrinsic objectives* predictably trade off between the two obstacles: curiosity- or novelty-driven exploration [90, 81, 30] amplifies sensitivity to states, yet visits otherwise unreachable regions, while surprise minimisation under the free-energy principle [34] acts as a stabiliser, yet may drive convergence to the hard-to-escape attractors that the multiplicity argument warns about [33]. *(iii) Gated adaptation* prevents systems from accumulating experience beyond a horizon over which prediction is reliable. *(iv) Circuit-breakers* pause or roll back learning when monitors detect loss of predictability or towards uncharacterised basins of attraction.

These actions are voluntary and thus do not fully overcome the obstacles. Whether developers deploy evaluable designs and embed runtime controls is ultimately their choice: market and reputational pressure may motivate some, but not all. Closing this gap will likely require connecting the recommendations above to existing and future governance regimes. We leave this to future work.

7 Alternative views

There is no way to design general-purpose CL systems which avoid chaoticity and multiplicity of attractors. While possible, this is hard to say without empirical evidence. The evidence we would gather by starting trajectory-centred evaluation now and by attempting to design systems amenable to evaluation (which we advocate) is necessary to confirm or dispute this alternative position.

It is already impossible to extensively test LLM agents pre-deployment, as novel architectures can be developed by users. It is true that advances are needed to characterise the behavioural profiles possible with different agent architectures for a given underlying LLM. Moreover, LLM agents show CL1 and could include CL2 or CL3. Thus, trajectory-centred evaluation is part of the practice needed to characterise behaviour of agents for a given LLM and this position is compatible with our own.

Existing monitoring tools (e.g., output filters) are sufficient for CL systems. Such monitors will still play a role to exclude, e.g., clearly harmful outputs. However, CL systems may develop new capabilities and propensities that output filters will miss or detect when it is already too late.

Landscape characterisation is not different from the goal of current evaluation practices, that include pre-deployment fine-tuning, red-teaming and elicitation to characterise possible behaviours. While the aim is analogous, current evaluation practices ignore specific (future) deployment information for more refined trajectory-based anticipation of benign and harmful uses.

We should rely on evaluation practices from traditional online learning. Traditional online learning focuses on a system sequentially learning well-defined tasks [55]. There, we can subject the system to comprehensive tests after each learning step. This is impossible for CL systems that are general-purpose and learn individually for each user, which expands the number of tasks and instances to test.

8 Conclusions

We are not the first to conclude that AI evaluation is not ready for CL systems: for medical devices, Vokinger et al. [98] argue for post-approval monitoring and running standardised tests while learning occurs, while recent surveys of self-evolving LLM agents [3, 32] stress the need of continuous evaluation. However, for general-purpose systems undergoing CL, extensively assessing every evolution step is infeasible. In this paper, we proposed a tractable alternative: trajectory-centred evaluation, which couples pre-deployment trajectory elicitation with post-deployment predictive monitors. Together, these characterise the landscape of behaviours a system can develop and forecast where a deployed system is heading. Trajectory-centred evaluation is most effective when layered with complementary defences: input/output filters for features not affected by learning (e.g., toxic language), transparency embedded in the evolution mechanisms (e.g., human-readable memory [47] or chain of thought [57]), and tracking broad indicators of deployed systems (as in pharmacovigilance, through reporting channels [27] and other indicators which may be affected by deployed systems [11]). Combined with trajectory-centred evaluation, these constitute a defence-in-depth strategy against CL systems developing undesirable behavioural profiles.

Acknowledgments and Disclosure of Funding

We acknowledge productive discussions with Sören Mindermann, Alan Cooney, Lisa Soder, Shalaleh Rismani, Marius Hobbhah, Joe Castellano, Nikola Jurkovic, Josh Tenenbaum, Thomas Kwa, Jacob Davies. We thank Robert Trager, Joost van de Weijer, Feifei Zhao, Yi Zeng, Pega Maham for their interest and support. Joe Castellano also provided strategic support.

LP is supported by Coefficient Giving (previously Open Philanthropy). SOh received funding from the Leverhulme Trust and Survival and Flourishing Fund. JHO's research is supported by OpenAI's grant to the 'AI Progress through the Lens of Predictable AI Ecosystems' programme, which is based at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. JHO also received funding by EUR2024-153548 (PREDAIT) "Towards Predictable AI" from "Spanish Europe Excelencia" 2024. This work has been supported by project CIPROM/2022/6 (FASSLOW), funded by Generalitat Valenciana; by the Spanish grant PID2024-162030OB-I00 (ROBIN), funded by MCIN/AEI/10.13039/501100011033 and by ERDF, A way of making Europe; by the VANTAGE project (Grant Agreement No. 101249800) funded by the European Union through the ECCC; and by the HIDDEN project (Grant Agreement No. 101202228) projects, funded by the European Union.

References

[1] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[2] Eitan Altman. *Constrained Markov decision processes*. Routledge, 1999.

[3] Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence, 2025.

[4] Anthropic. Agent skills overview. Claude Documentation, 2024. Accessed: 2025.

[5] Anthropic. Claude introduces memory for teams at work. https://www.anthropic.com/news/memory, 2025. Accessed: 5th May 2026.

[6] Ludwig Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer, Berlin, Heidelberg, 1998.

[7] Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*, 2024.

[8] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[9] Fazl Barez. When AI systems learn during deployment, our safety evaluations break. Oxford Martin AI Governance Initiative Blog, January 2026. Accessed: 2026-05-18.

[10] Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Nicolaus Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.

[11] Rachel E Behrman, Joshua S Benner, Jeffrey S Brown, Mark McClellan, Janet Woodcock, and Richard Platt. Developing the sentinel system—a national resource for evidence development. *New England Journal of Medicine*, 364(6):498–499, 2011.

[12] Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. *arXiv preprint arXiv:2512.24695*, 2025.

[13] Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[14] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models, 2024.

[15] Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

[16] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko,

Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[18] John Burden, Marko Tešić, Lorenzo Pacchiardi, and José Hernández-Orallo. Paradigms of AI evaluation: Mapping goals, methodologies and culture, 2025.

[19] Ryan Burnell, Yumeya Yamamori, Orhan Firat, Kate Olszewska, Steph Hughes-Fitt, Oran Kelly, Isaac R. Galatzer-Levy, Meredith Ringel Morris, Allan Dafoe, Alison M. Snyder, Noah D. Goodman, et al. Measuring progress toward AGI: A cognitive taxonomy. Technical report, Google DeepMind, March 2026.

[20] David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. Beyond prompts: Dynamic conversational benchmarking of large language models, 2024.

[21] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[22] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.

[23] Cheng Chen, Maria D Molina, Mengqi Liao, and Eugene Cho Snyder. Relational gains, privacy strains: Exploring users' perceptions and experiences with chatgpt's memory feature. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2026.

[24] Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning*. Morgan and Claypool, 2nd edition, 2018.

[25] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

[26] Leono Chua, Motomasa Komuro, and Takashi Matsumoto. The double scroll family. *IEEE transactions on circuits and systems*, 33(11):1072–1118, 1986.

[27] Valeri Craigle. Medwatch: The fda safety information and adverse event reporting program. *Journal of the Medical Library Association*, 95(2):224, 2007.

[28] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, 2014.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] Emmanuel Dupoux, Yann LeCun, and Jitendra Malik. Why ai systems don't learn and what to do about it: Lessons on autonomous learning from cognitive science. *arXiv preprint arXiv:2603.15381*, 2026.

[31] Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, et al. Cartridges: Lightweight and general-purpose long context representations via self-study. *arXiv preprint arXiv:2506.06266*, 2025.

[32] Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems, 2025.

[33] Zafeirios Fountas, Noor Sajid, Pedro Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11662–11675. Curran Associates, Inc., 2020.

[34] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[35] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

[36] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022.

[37] Tom Gibbs, Ethan Kosak-Hine, George Ingebretsen, Jason Zhang, Julius Broomfield, Sara Pieri, Reihaneh Iranmanesh, Reihaneh Rabbany, and Kellin Pelrine. Emerging vulnerabilities in frontier models: Multi-turn jailbreak attacks, 2024.

[38] Anmol Goel, Cornelius Emde, Sangdoo Yun, Seong Joon Oh, and Martin Gubri. Privacy collapse: Benign fine-tuning can break contextual privacy in language models, 2026.

[39] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Multi-token attention. *arXiv preprint arXiv:2504.00927*, 2025.

[40] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey, 2023.

[41] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[42] Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. A definition of AGI, 2025.

[43] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[44] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[45] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[46] Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions, 2026.

[47] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. Memory sandbox: Transparent and interactive memory management for conversational agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3, 2023.

[48] Jonas Hübotter, Leander Diaz-Bone, Ido Hakimi, Andreas Krause, and Moritz Hardt. Learning on the job: Test-time curricula for targeted reinforcement learning. *arXiv preprint arXiv:2510.04786*, 2025.

[49] Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

[50] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[51] Olli Järviniemi, Oliver Makins, Jacob Merizian, Robert Kirk, and Ben Millwood. Propensity inference: Environmental contributors to llm behaviour, 2026.

[52] Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024.

[53] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[54] Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why fine-tuning encourages hallucinations and how to fix it, 2026.

[55] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives, 2022.

[56] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[57] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

[58] Vojtech Kovarík, Eric Olav Chen, Sami Petersen, Alexis Ghersengorin, and Vincent Conitzer. AI testing should account for sophisticated strategic behaviour. *arXiv preprint arXiv:2508.14927*, 2025.

[59] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation, 2025.

[60] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[61] Hanyu Li, Haoyu Liu, Tingyu Zhu, Tianyu Guo, Zeyu Zheng, Xiaotie Deng, and Michael I Jordan. Ida-bench: Evaluating llms on interactive guided data analysis. *arXiv preprint arXiv:2505.18223*, 2025.

[62] Haoxuan Li, Mingyu Derek Ma, Jen-tse Huang, Zhaotian Weng, Wei Wang, and Jieyu Zhao. Biasinspector: Detecting bias in structured data through llm agents. *arXiv preprint arXiv:2504.04855*, 2025.

[63] Jinyang Li, Nan Huo, Yan Gao, Jiayi Shi, Yingxiu Zhao, Ge Qu, Bowen Qin, Yurong Wu, Xiaodong Li, Chenhao Ma, Jian-Guang Lou, and Reynold Cheng. Are large language models ready for multi-turn tabular data analysis? In *Forty-second International Conference on Machine Learning*, 2025.

[64] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[65] Chengyuan Liu, Shihang Wang, Yangyang Kang, Lizhi Qing, Fubang Zhao, Changlong Sun, Kun Kuang, and Fei Wu. More than catastrophic forgetting: Integrating general capabilities for domain-specific LLMs. *arXiv preprint arXiv:2405.17830*, 2024.

[66] Winfried Lohmiller and Jean-Jacques E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[67] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[68] Kenji Matsumoto and Ichiro Tsuda. Noise-induced order. *Journal of Statistical Physics*, 31(1):87–106, 1983.

[69] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[70] Steven W McDonald, Celso Grebogi, Edward Ott, and James A Yorke. Fractal basin boundaries. *Physica D: Nonlinear Phenomena*, 17(2):125–153, 1985.

[71] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024.

[72] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[73] Terrie E. Moffitt, Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent W. Roberts, Stephen Ross, Malcolm R. Sears, W. Murray Thomson, and Avshalom Caspi. A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7):2693–2698, 2011.

[74] Meredith Ringel Morris, Dan Altman, Haydn Belfield, Arthur Goemans, Hasan Iqbal, Ryan Burnell, Iason Gabriel, Samuel Albanie, and Allan Dafoe. Characterizing model jaggedness supports safety and usability. 2026.

[75] Siddharth Narayanan, James D. Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Aviary: training language agents on challenging scientific tasks, 2024.

[76] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. Mlgym: A new framework and benchmark for advancing ai research agents, 2025.

[77] Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.

[78] OpenAI. Memory and new controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/, 2024. Accessed: 5th May 2026.

[79] OpenAI. Where the goblins came from. https://openai.com/index/where-the-goblins-came-from/, April 2026. Accessed: 2026-05-01.

[80] Edward Ott, Celso Grebogi, and James A. Yorke. Controlling chaos. *Physical Review Letters*, 64(11):1196–1199, 1990.

[81] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017.

[82] Jonathan Prunty, Aoife O'Flynn, Patrick Quinn, and Lucy G Cheke. Intuit: Investigating intuitive reasoning in humans and language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[83] Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 431–445, Singapore, December 2023. Association for Computational Linguistics.

[84] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[85] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345, 2007.

[86] Daniel Romero-Alvarado, Fernando Martínez-Plumed, Lorenzo Pacchiardi, Hugo Save, Siddhesh Milind Pawar, Behzad Mehrbakhsh, Pablo Antonio Moreno Casares, Ben Slater, Paolo Bova, Peter Romero, et al. Capabilities ain't all you need: Measuring propensities in ai. *arXiv preprint arXiv:2602.18182*, 2026.

[87] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025.

[88] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

[89] Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206(R), 2003.

[90] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation. *IEEE Trans. on Auton. Ment. Dev.*, 2(3):230–247, September 2010.

[91] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

[92] Adi Simhi, Fazl Barez, Martin Tutek, Yonatan Belinkov, and Shay B. Cohen. Old habits die hard: How conversational history geometrically traps LLMs, 2026.

[93] Shezheng Song, Hao Xu, Jun Ma, Shasha Li, Long Peng, Qian Wan, Xiaodong Liu, and Jie Yu. How to alleviate catastrophic forgetting in LLMs finetuning? hierarchical layer-wise and element-wise regularization. *arXiv preprint arXiv:2501.13669*, 2025.

[94] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

[95] Yuanyuan Sun, Timothy Parker, Lara Gierschmann, Sana Shams, Teo Canmetin, Mathieu Duteil, Rokas Gipiškis, and Ze Shen Chin. Defining ai models and ai systems: A framework to resolve the boundary problem, 2026.

[96] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny

Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed,

Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews,

CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[97] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[98] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. Continual learning in medical devices: Fda's action plan and beyond. *The Lancet Digital Health*, 3(6):e337–e338, 2021.

[99] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. arXiv:2302.00487.

[100] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[101] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents, 2026.

[102] Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. *arXiv preprint arXiv:2603.19461*, 2026.

[103] Jiayi Zhang, Yiran Peng, Fanqi Kong, Cheng Yang, Yifan Wu, Zhaoyang Yu, Jinyu Xiang, Jianhao Ruan, Jinlin Wang, Maojia Song, HongZhang Liu, Xiangru Tang, Bang Liu, Chenglin Wu, and Yuyu Luo. Autoenv: Automated environments for measuring cross-environment agent learning, 2025.

[104] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[105] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

[106] Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. LifelongAgentBench: Evaluating LLM agents as lifelong learners, 2025.

[107] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, et al. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652(8108):58–67, 2026.