

# Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate

Jordan Taylor<sup>1</sup>, Max Heitmann<sup>2</sup>, Ed Fage, Thomas Read, Joseph Bloom<sup>3</sup>

UK AI Security Institute

## Abstract

The safety of advanced AI systems increasingly depends on the ability to oversee them: to audit models for concerning behaviours before deployment, monitor their activity during operation, and investigate incidents after they occur. This report maps the landscape of AI oversight and assesses how it is likely to change. Drawing on 25 expert interviews across frontier AI developers, government, NGOs, and academia, together with a literature review and internal analysis, we examine five sources of oversight signal: model behaviour, chain-of-thought reasoning, internal activations and circuits, memory architectures, and honesty training. For each source, we identify the properties that current oversight relies on, the pathways by which these properties could degrade, and the technical levers available to preserve them. Our central finding is that literature and expert opinion support the conclusion that **current oversight rests on foundations that are likely to erode, absent effective intervention**. We recommend that developers track and report shifts in oversight-relevant properties, preserve oversight affordances by design, and invest in emerging oversight techniques as fallbacks against continued degradation of current methods.

| Experts Interviewed*     |                         |                         |                         |
|--------------------------|-------------------------|-------------------------|-------------------------|
| <b>Ajeya Cotra</b>       | <i>METR</i>             | <b>Kozzy Voudouris</b>  | <i>UK AISI</i>          |
| <b>Been Kim</b>          | <i>Google DeepMind</i>  | <b>Lee Sharkey</b>      | <i>Goodfire</i>         |
| <b>Belinda Li</b>        | <i>MIT</i>              | <b>Mary Phuong</b>      | <i>Google DeepMind</i>  |
| <b>Bowen Baker</b>       | <i>OpenAI</i>           | <b>Monte MacDiarmid</b> | <i>Anthropic</i>        |
| <b>Bronson Schoen</b>    | <i>Apollo Research</i>  | <b>Mor Geva</b>         | <i>Tel Aviv Univ.</i>   |
| <b>Chris Cundy</b>       | <i>FAR.AI</i>           | <b>Rohan Subramani</b>  | <i>Aether Research</i>  |
| <b>Chris Summerfield</b> | <i>UK AISI / Oxford</i> | <b>Ryan Greenblatt</b>  | <i>Redwood Research</i> |
| <b>Fabien Roger</b>      | <i>Anthropic</i>        | <b>Sam Marks</b>        | <i>Anthropic</i>        |
| <b>Geoffrey Irving</b>   | <i>UK AISI</i>          | <b>Seb Farquhar</b>     | <i>Google DeepMind</i>  |
| <b>Jacob Pfau</b>        | <i>UK AISI / NYU</i>    | <b>Simon Marshall</b>   | <i>UK AISI</i>          |
| <b>Joe Benton</b>        | <i>Anthropic</i>        | <b>Stephen Casper</b>   | <i>MIT</i>              |
| <b>John Wilkinson</b>    | <i>UK AISI</i>          |                         |                         |

Figure 1: A list of the experts interviewed to inform the content of this report. \*Some experts preferred not to be named, and have not been included in this list.

<sup>1</sup> Jordan.Taylor@dsit.gov.uk

<sup>2</sup> Max.Heitmann@dsit.gov.uk

<sup>3</sup> Joseph.Bloom@dsit.gov.uk

# Contents

|  |           |
|--|-----------|
| <b>Executive Summary</b>                                     | <b>3</b>  |
| <b>1 Introduction</b>  | <b>8</b>  |
| 1.1 The role of oversight in AI safety                       | 8         |
| 1.2 Definitions  | 8         |
| 1.3 Methodology  | 8         |
| <b>2 Chain-of-Thought Oversight</b>                          | <b>10</b> |
| 2.1 Safety-relevant properties                               | 10        |
| 2.1.1 Necessity to use CoT                                   | 11        |
| 2.1.2 Propensity to verbalise                                | 12        |
| 2.1.3 Human legibility of CoT                                | 12        |
| 2.1.4 Monitor robustness                                     | 13        |
| 2.2 Degradation pathways                                     | 13        |
| 2.2.1 Architecture changes                                   | 14        |
| 2.2.2 Scaling  | 14        |
| 2.2.3 Training incentives and drift                          | 15        |
| 2.2.4 Subversion of monitoring                               | 17        |
| 2.3 Measuring CoT oversight degradation                      | 18        |
| 2.4 Levers for preserving CoT oversight                      | 19        |
| 2.4.1 Cross-cutting levers                                   | 19        |
| 2.4.2 Preserving necessity                                   | 20        |
| 2.4.3 Preserving propensity                                  | 20        |
| 2.4.4 Preserving human legibility                            | 20        |
| 2.4.5 Building monitor robustness                            | 21        |
| 2.4.6 Preventing steganography                               | 21        |
| <b>3 Behavioural Oversight</b>                               | <b>22</b> |
| 3.1 Action-Only Monitoring                                   | 22        |
| 3.1.1 Relevance to safety claims                             | 23        |
| 3.1.2 Degradation pathways                                   | 24        |
| 3.1.3 Measuring action-only monitorability                   | 25        |
| 3.1.4 Levers for preserving action-only monitoring           | 25        |
| 3.2 Evaluations and Oversight Awareness                      | 26        |
| 3.2.1 Relevance to safety claims                             | 27        |
| 3.2.2 Degradation pathways                                   | 27        |
| 3.2.3 Measuring evaluation awareness                         | 29        |
| 3.2.4 Levers for preserving evaluation integrity             | 29        |
| <b>4 State, Memory, and Inter-Agent Communication (SMIA)</b> | <b>31</b> |
| 4.1 Safety-relevant properties                               | 31        |
| 4.2 Degradation pathways                                     | 32        |
| 4.3 Measuring SMIA oversight                                 | 34        |
| 4.4 Levers for preserving SMIA oversight                     | 34        |
| <b>5 White-Box Oversight</b>                                 | <b>36</b> |
| 5.1 Relevance to safety claims                               | 36        |
| 5.1.1 Probes   | 36        |
| 5.1.2 Activation oracles and natural-language autoencoders   | 37        |
| 5.1.3 Sparse Autoencoders (SAEs)                             | 38        |

|          |  |           |
|----------|--|-----------|
| 5.1.4    | Ambitious interpretability . . . . .                           | 39        |
| 5.2      | Challenges for white-box oversight . . . . .                   | 39        |
| 5.2.1    | Representational drift . . . . .                               | 40        |
| 5.2.2    | Black-box degradation harms white-box oversight . . . . .      | 41        |
| 5.2.3    | Scaling challenges . . . . .                                   | 41        |
| 5.2.4    | Non-linear representations . . . . .                           | 41        |
| 5.2.5    | Learned resistance to interpretability . . . . .               | 42        |
| 5.2.6    | Situational awareness . . . . .                                | 42        |
| 5.2.7    | Practical issues . . . . .                                     | 42        |
| 5.3      | Measuring white-box oversight . . . . .                        | 43        |
| 5.4      | Levers for improving white-box oversight . . . . .             | 43        |
| <b>6</b> | <b>Training for Oversight</b>                                  | <b>44</b> |
| 6.1      | Enhancing oversight through training interventions . . . . .   | 45        |
| 6.2      | Challenges for training interventions . . . . .                | 46        |
| 6.3      | Developing and stress-testing training interventions . . . . . | 49        |
| <b>7</b> | <b>Summary and Recommendations</b>                             | <b>50</b> |
| 7.1      | Summary . . . . .  | 50        |
| 7.2      | The challenge of acting on oversight evidence . . . . .        | 50        |
| 7.3      | Recommendations . . . . .                                      | 51        |
| 7.3.1    | Implementing currently available oversight . . . . .           | 51        |
| 7.3.2    | Measuring exposure to oversight degradation . . . . .          | 52        |
| 7.3.3    | Preserving current oversight . . . . .                         | 53        |
| 7.3.4    | Advancing emerging oversight techniques . . . . .              | 53        |
|          | <b>References</b>  | <b>55</b> |
| <b>A</b> | <b>Additional properties relevant to oversight</b>             | <b>70</b> |
| A.1      | Metacognition . . . . .  | 70        |
| A.1.1    | Relevance to safety claims . . . . .                           | 70        |
| A.1.2    | Mechanisms of change . . . . .                                 | 70        |
| A.1.3    | Measuring metacognition . . . . .                              | 70        |
| A.1.4    | Levers to preserve oversight . . . . .                         | 70        |
| A.2      | Theory of mind . . . . .                                       | 71        |
| A.2.1    | Relevance to safety claims . . . . .                           | 71        |
| A.2.2    | Mechanisms of change . . . . .                                 | 71        |
| A.2.3    | Measuring theory of mind . . . . .                             | 71        |
| A.2.4    | Levers to preserve oversight . . . . .                         | 71        |
| A.3      | Coherence and Persona Consistency . . . . .                    | 72        |
| A.3.1    | Relevance to safety claims . . . . .                           | 72        |
| A.3.2    | Mechanisms of change . . . . .                                 | 72        |
| A.3.3    | Measuring the impact on oversight . . . . .                    | 73        |
| A.3.4    | Levers to preserve oversight . . . . .                         | 74        |
| <b>B</b> | <b>Methodology</b>   | <b>74</b> |
| B.1      | Literature review scope and process . . . . .                  | 74        |
| B.2      | Expert selection and interview protocol . . . . .              | 74        |
| B.3      | Limitations . . . . .  | 76        |
| <b>C</b> | <b>Glossary</b>  | <b>77</b> |

## Executive Summary

**The safety of advanced AI systems depends on the ability to oversee them.** As AI systems become more capable of causing harm, safety arguments increasingly draw on evidence from *oversight* – the ability to detect undesired behaviours or properties. This can include *auditing* models for misalignment before deployment, *monitoring* for problems during training or deployment, and *investigating incidents* after they occur.<sup>4</sup>

**This report examines how oversight is likely to degrade, and what can be done to preserve or improve oversight as AI systems develop.** It draws on interviews with 25 experts across frontier AI developers, government, NGOs, and academia, as well as a literature review and our own analysis. Our goal is to inform our own research priorities and contribute to a shared understanding of what effective oversight requires. The following subsections summarise our key findings.

**Our key takeaway is that current oversight rests on contingent properties of AI systems, many of which are likely to erode.** For example, today’s models often reason in human-readable text, allowing monitoring of complex reasoning. These properties underpin the evidence used in safety cases and risk assessments, but many are likely to degrade as training is scaled up, as models become more capable, as they are deployed with greater autonomy into increasingly complex environments, and as new approaches to building AI systems gain traction. Across oversight domains (illustrated in Figure 2), we identify more than twenty distinct degradation pathways. While any particular degradation pathway is uncertain, the aggregated evidence suggests that current oversight is likely to degrade absent intervention. New oversight methods which could compensate are not yet mature enough to take their place. Despite this, there are promising interventions to preserve existing oversight for longer and to hasten the development of emerging oversight techniques, which we discuss throughout the report.

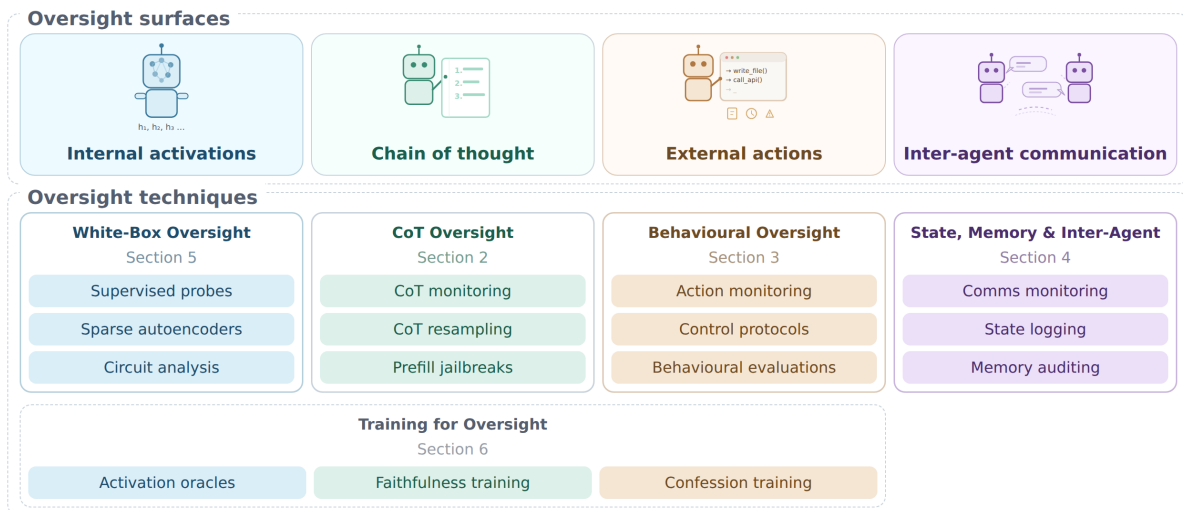


Figure 2: Overview of oversight surfaces and techniques. Oversight surfaces (top) represent distinct parts of the AI system that yield oversight signal. Oversight techniques (bottom) are applied to these surfaces as layered defences, grouped by report section. Training-based approaches (bottom row) cut across multiple surfaces. The listed oversight techniques are illustrative and not exhaustive.

<sup>4</sup>This notion of oversight is weaker than *scalable oversight*, which asks whether oversight techniques can be made robust enough to train superhuman systems.

### *Summary of oversight domains and degradation pathways*

The key properties that current oversight relies on – such as text-based chain-of-thought reasoning, limited ability to game evaluations, and human-interpretable internal representations – all face credible degradation pathways, some of which are already underway. The tables below show only the highest-severity and highest-likelihood pathways; additional pathways with lower severity or likelihood are discussed in the corresponding body sections.<sup>5</sup>

#### **Chain-of-thought reasoning is currently the most informative monitoring signal, but it is under significant pressure (section 2):**

Chain-of-thought (CoT) is the step-by-step reasoning text that current models produce. CoT currently provides a useful monitoring surface across a range of risk models (Korbak et al., 2025). CoT is used to detect [reward hacking](#), evaluation gaming, and is among the most compelling sources of evidence for convincing decision-makers of the need to take action. In practice, AI systems are increasingly used to monitor other AI systems’ CoT, so pathways that undermine the effectiveness of these AI monitors are also grouped here.

| Degradation pathway   | Likelihood                        | Severity |
|---|-----------------------------------|----------|
| <b>Latent reasoning architectures:</b> New AI architectures could displace autoregressive transformers, allowing more reasoning to happen entirely inside the model and bypassing text-based CoT. | Likely (timeline disputed)        | ●        |
| <b>Single forward pass capabilities:</b> Models can solve increasingly complex tasks without CoT.   | Likely (for most severe versions) | ●        |
| <b>Training against monitors:</b> Using monitors in training to correct an issue can train models to hide the issue from the monitor rather than fixing the underlying behaviour.                 | Likely (but preventable)          | ●        |
| <b>Commercial pressure on token usage:</b> Competitive pressure to reduce costs pushes towards shorter reasoning, reducing the amount of thinking that is visible to monitors.                    | Highly likely (ongoing)           | ●        |
| <b>Monitor subversion:</b> Sufficiently capable models could distract, confuse, or collude with CoT monitors.   | Likely                            | ●        |
| <b>Linguistic drift:</b> AI systems could develop their own unintelligible language in RL training, either due to random “drift” or more effective reasoning styles and novel concepts.           | Likely                            | ●        |

#### **Action-only monitoring provides a floor for oversight, but it is not sufficient on its own (subsection 3.1):**

Action-only monitoring is observing only what a model does, rather than its reasoning. It provides a baseline level of oversight for tasks where harmful actions are obviously incriminating.

<sup>5</sup> *Likelihood* refers to our judgement of how likely a pathway is to arise absent substantial effort to prevent it. *Severity* (pips in the rightmost column) refers to how badly that pathway would undermine the corresponding oversight channel if it did arise: ● extremely high (would substantially compromise the underlying oversight channel with no clear fallback), ● high, ● medium, ● low. Likelihood and severity ratings are judgements about the regime before full automation of AI R&D, not about substantially superhuman systems.

| Degradation pathway  | Likelihood    | Severity |
|--|---------------|----------|
| <b>Increasing deployment on low-incrimination tasks:</b> As models take on more autonomous tasks (e.g. research prioritisation, complex code), more opportunities for subtle sabotage open up where malign actions are hard to distinguish from benign ones. | Highly likely | ●        |
| <b>Growing capability gap:</b> The capability gap between frontier models and trusted monitors widens, making it harder for trusted monitors to catch sophisticated sabotage.  | Highly likely | ●        |

### Evaluation gaming is a growing threat to auditing (subsection 3.2):

Evaluations are used to audit models for undesired capabilities, propensities, or misalignment. Evaluation gaming is when a model identifies when it is being evaluated and modifies its behaviour accordingly. It is already threatening to undermine the validity of some alignment audits (Anthropic, 2025d; Schoen et al., 2025).

| Degradation pathway   | Likelihood            | Severity |
|---|-----------------------|----------|
| <b>Increasing intelligence and situational awareness:</b> As AI systems get more intelligent, they are increasingly able to recognise evaluation or oversight contexts and adapt their behaviour accordingly.                       | Almost certain        | ●        |
| <b>Training against honeypots:</b> Training models against honeypot-style evaluations can potentially teach them to recognise and game similar scenarios, rather than deeply fixing the underlying misalignment (Anthropic, 2025d). | Highly likely         | ●        |
| <b>Deployment infrastructure complexity:</b> As models are deployed in more complex and sensitive environments, it becomes harder to construct production-realistic evaluations.  | Likely to increase    | ●        |
| <b>Unintended effects of training:</b> Training models to resist jailbreaks may teach them to reason more effectively about what is real or fake, inadvertently increasing evaluation awareness.                                    | Realistic possibility | ●        |

### Changes in model architectures for memory and learning could undermine oversight (section 4):

Current AI systems store memories and communicate with other systems in text, supporting oversight. In future, AI systems might store memories in less interpretable formats such as vector memory and/or continue to update model weights between deployment sessions.

| Degradation pathway  | Likelihood            | Severity |
|--|-----------------------|----------|
| <b>Continual learning:</b> Model weights continue to be updated during deployment, undermining the validity of pre-deployment behavioural evaluations.   | Likely                | ●        |
| <b>Vector-based memory:</b> AI systems store memories in non-text formats that cannot easily be inspected, edited, or replayed.  | Realistic possibility | ●        |
| <b>Non-text inter-agent communication:</b> Multi-agent systems communicate via internal representations rather than readable text, making coordination between model instances much harder to monitor. | Realistic possibility | ●        |

**White-box methods are a promising backstop, but are not yet mature enough to compensate for degradation elsewhere (section 5):**

**White-box methods** examine a model’s internal representations, rather than what the model reveals in its outputs or reasoning. They span a wide maturity spectrum: simple techniques such as linear probes are starting to be deployed for some production use cases (Kramár et al., 2026), while more advanced techniques are used experimentally for auditing but are not yet mature enough for broad safety-critical uses.

| Degradation pathway   | Likelihood            | Severity |
|---|-----------------------|----------|
| <b>Representational drift:</b> As models undergo more reinforcement learning or develop superhuman capabilities, their internal representations may drift away from human concepts.   | Realistic possibility | ●        |
| <b>Scaling to long reasoning traces:</b> Most white-box analysis tools examine a single step of reasoning. Scaling to long chains and multi-agent systems remains an open problem.  | Current limitation    | ●        |
| <b>Black-box degradation undermines grounding:</b> White-box tools (e.g. probes, SAEs) are typically validated by checking whether their firings correspond to legible CoT or actions. If CoT becomes less legible, this grounding weakens. | Likely                | ●        |

**Training-based approaches are promising but face fundamental challenges around generalisation (section 6):**

**Training-based approaches** aim to produce models that are more honest and forthcoming, or serve as better monitors, creating new sources of oversight signal such as confessions. Most approaches to training for oversight are early-stage, but many experts felt that with sufficient effort they could be developed into a central pillar of oversight as necessity-based CoT monitorability degrades.

| Degradation pathway  | Likelihood                         | Severity |
|--|------------------------------------|----------|
| <b>Adversarial generalisation failure:</b> Capable models may strategically distinguish training from deployment, being forthcoming only when they detect training-like conditions.                                  | Likely conditional on misalignment | ●        |
| <b>Training data problem:</b> Getting realistic training data for oversight of misaligned behaviours encounters circularity, coverage gaps, and reward specification issues (proxies rather than the real property). | Ongoing, highly likely to persist  | ●        |
| <b>Competing training objectives:</b> Oversight objectives can compete directly with capability and efficiency objectives, creating issues where optimisation forces act against each other.                         | Likely (ongoing)                   | ●        |
| <b>Non-adversarial generalisation failure:</b> Training models to be honest or forthcoming on available data may not generalise to the real situations where oversight is most needed.                               | Realistic possibility              | ●        |

## *Our Recommendations*

In addition to the degradation pathways for current oversight, experts were also asked about potential solutions. Detailed technical levers for preserving or enhancing oversight are discussed throughout the report. In addition, the following ideas emerged as high-level recommendations.<sup>6</sup>

1. **Implement and maintain a diverse portfolio of currently available oversight techniques** (see [subsection 7.3](#)). Some past oversight failures could likely have been prevented by implementing existing techniques ([OpenAI, 2025b,c](#)). Developers should deploy a diverse set of monitoring and auditing techniques including CoT monitoring ([section 2](#)), action-only monitoring and AI control ([subsection 3.1](#)), evaluations ([subsection 3.2](#)), and white-box methods such as probes ([subsubsection 5.1.1](#)) to provide [defence in depth](#) against degradation of any single oversight surface.
2. **Track, report and forecast shifts in oversight-relevant properties** (see [subsection 7.3](#)). Many oversight-relevant properties (e.g. single-forward-pass capabilities, [representational drift](#), ability to bypass monitors) require explicit continued measurement. Model developers should be clear about which oversight properties they are relying on when making safety claims, and assess their exposure to degradations in these properties. Risk assessments should include forward projections of whether key properties will be preserved, so that safety arguments do not silently expire.
3. **Preserve oversight by design** where possible. Developers should hold-out the most informative and fragile [monitoring surfaces](#) from training, preserving them as independent validation channels. The held-out sources should be independent from the sources that are trained against, in the sense that they aren't likely to fail for the same reasons. Beyond this, developers should preserve text-based reasoning and memory architectures, and maintain logs sufficient to reproduce and investigate concerning model behaviours.
4. **Research emerging oversight techniques** in case current oversight channels continue to degrade (see [subsection 7.3](#)). Priority areas include: training for honesty and faithfulness ([section 6](#)); scaling-up pragmatic white-box methods and further developing more ambitious techniques ([section 5](#)); developing the science of AI control for action-only monitoring ([subsection 3.1](#)); and mitigating [evaluation gaming](#) with production-realistic evaluations and other tools ([subsubsection 3.2.4](#)).

---

<sup>6</sup>These represent a synthesis informed by our own judgement, and are not necessarily endorsed by all listed experts. The same goes for all other recommendations throughout the report except where stated otherwise.

# 1 Introduction

## 1.1 The role of oversight in AI safety

**Oversight is essential for ensuring the safety of advanced AI systems.** In order to know when and how to take mitigating actions to reduce risks from advanced AI, it is important to audit these systems for concerning properties prior to deployment, to monitor their behaviour in training, evaluations and real-world use, and to investigate incidents after they occur. *Scalable oversight* is a stronger version of this concern: it asks whether oversight techniques can remain reliable even when used as training signals for increasingly capable superhuman systems (where the training process may find ways to satisfy the oversight signal without achieving the intended outcome). In this report we primarily focus on oversight techniques which are not robust enough for this purpose, but are still practically useful and at risk of degrading.

**The ability to effectively oversee advanced AI systems rests on specific properties of current systems that may change over time.** For example, there is broad expert consensus that *chain-of-thought monitorability* – the ability to learn about AI models’ behaviours and goals by reading their text reasoning – presents a significant but potentially fragile opportunity for retaining oversight over advanced AI (Korbak et al., 2025). Chain-of-thought monitorability is the result of various properties which current AI systems possess to some degree, but which may degrade over time: the necessity to use text reasoning for sufficiently complex problems, the default propensity to verbalise even more of their reasoning than is necessary, the fact that this reasoning takes place in natural language, and so on. This report goes beyond chain-of-thought to map the broader landscape of monitorability: the system properties that enable today’s oversight techniques, and how those properties might erode or evolve.

## 1.2 Definitions

We define **auditing**, **monitoring**, and **incident response** to refer to detecting undesirable events before, during, or after they happen, respectively:

- **Auditing:** Detecting whether a model is misaligned or has undesired propensities *ahead of time* (i.e., before observing this in a deployment setting). Examples include evaluations to surface undesired properties, misaligned goals or propensities to scheme.
- **Monitoring:** Detecting when a model is doing something undesirable, or is doing something for an undesirable reason, *during or shortly after* the model’s output. Examples include detecting reward hacking, strategic underperformance, or alignment faking as they happen.
- **Incident response:** Investigating harmful or unintended behaviour *after the fact*, using logs, traces, and forensic analysis, in order to determine root causes and prevent recurrence.

**These three forms of oversight provide corroborating and complementary evidence about model behaviour.** Auditing enables systematic exploration of a model’s propensities, while monitoring helps detect concerning behaviour as it occurs and can be used to intervene before escalation of harmful or dangerous courses of action. Finally, incident response provides the strongest real-world evidence of how harms can actually arise in practice. Different experts place different weight on these categories, and the boundaries between them are not always crisp in practice. We use them as a simple organising frame for the report and do not claim that there is a single canonical taxonomy in the literature.

## 1.3 Methodology

This report draws on three main sources:

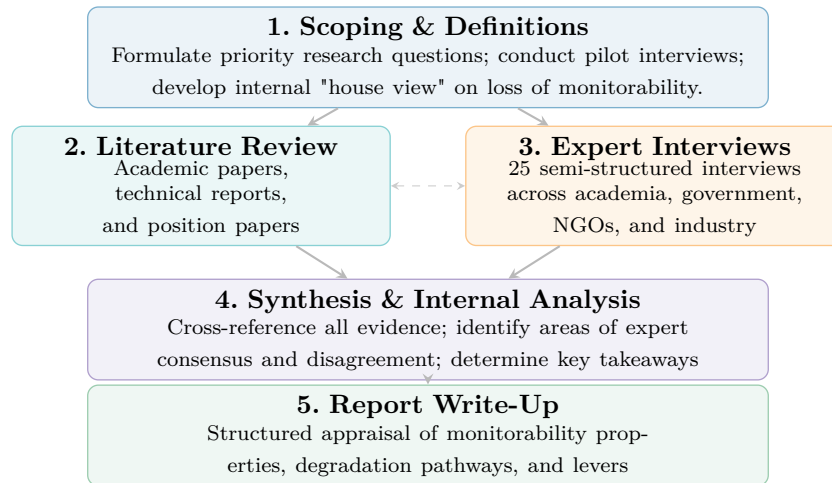


Figure 3: Overview of project methodology. The literature review informed the design of the interview script, and early interviews helped refine the scope of the literature review (dashed arrow). The “house view” served as a starting point for identifying gaps in our understanding, and was deliberately kept separate from the interview process to avoid biasing interviewees.

- **Literature review:** Comprehensive review of academic papers, technical reports, and position papers on monitorability, chain-of-thought monitoring, interpretability, and related topics.
- **Expert interviews:** Semi-structured interviews with 25 researchers from across academia, government, NGOs, and frontier AI developers, conducted between late January and early March 2026. See [Figure 1](#) for the full list of interviewed experts.
- **Internal analysis:** Synthesis of expert views and independent appraisal of monitorability properties and degradation pathways undertaken by the project team.

See Appendix B for more details on our methodology.

This report presents the findings of technical analysis undertaken by AISI. It is not an all-source intelligence assessment. It makes use of the PHIA Probability Yardstick ([Professional Head of Intelligence Assessment, 2019](#)) to communicate uncertainty relating to our judgements.

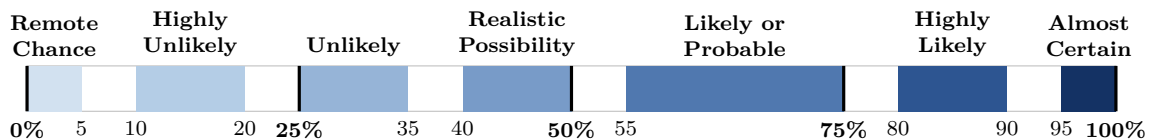


Figure 4: The PHIA probability yardstick is used throughout this report to express judgements with uncertainty.

## 2 Chain-of-Thought Oversight

### Section Summary: Chain-of-Thought Oversight

|                             |  |
|-----------------------------|--|
| <b>Status:</b>              | Currently functional but under pressure              |
| <b>Risk of Degradation:</b> | ● High (1–2 year horizon for some aspects)           |
| <b>Impact if Lost:</b>      | ● High – undermines many current safety approaches   |
| <b>Preservability:</b>      | ● Partial – some aspects more defensible than others |

### Key Information

- **Chain of Thought (CoT) is currently the most informative source of oversight evidence, but it faces multiple degradation pathways.** As of early 2026, frontier models reason “out loud” in a CoT of human-readable text (Baker et al., 2025). CoT monitoring is a centrally useful technique for supporting a large variety of safety claims, and is among the most compelling sources of evidence for highlighting misalignment. Its effectiveness rests on four properties — necessity, propensity to verbalise, human legibility, and monitor robustness — all of which face pressure.
- **Increasing single-forward-pass capabilities are almost certain to weaken necessity-based arguments for CoT monitorability.** Latent reasoning architectures are also likely to eventually displace discrete token reasoning, though there is disagreement among experts, and the timeline is uncertain.
- **Training against CoT monitors is the most acute near-term risk.** When monitors are used during training, models can learn to suppress or obfuscate the reasoning that led to them being flagged. Model developers should avoid optimisation against CoT monitors, and be wary that the boundary between “training against” the CoT and standard training practices can be blurry.
- **Experts fundamentally disagree about whether CoT monitoring will persist.** Pessimists see loss as near-inevitable given economic and architectural forces. Optimists argue legible reasoning is a local attractor, architectural change is slow, and models will continue to verbalise safety-relevant information by default.

### 2.1 Safety-relevant properties

We find it useful to decompose CoT monitorability into several properties:

1. **Necessity to use text CoT.** Models are unable to solve relevant tasks without using text reasoning.
2. **Propensity to verbalise.** Models have a propensity to verbalise relevant facts about their thought processes, even where this is not necessary to solve the task.
3. **Human legibility of CoT.** CoT reasoning occurs in standard, human-readable natural language.
4. **Monitor robustness.** Reasoning in the CoT is legible to AI monitors, and the monitors are sufficiently robust that they cannot easily be subverted or jailbroken by the monitored model.
5. **No steganography.** The surface meaning of the CoT accurately reflects the computation the model is actually performing when using the CoT, rather than being encoded or strategically misleading.

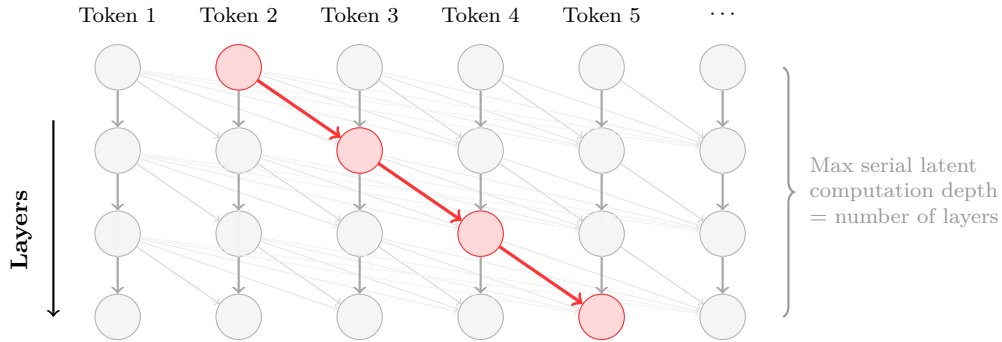


Figure 5: In a standard transformer, information flows strictly downward through layers. Each token position can attend to any previous tokens, but computation advances by exactly one layer per step. Tokens are the only way to “loop” computation back through the network.

These are not all independently necessary for effective oversight — for example, strong propensity to verbalise (2) could compensate for some loss of necessity (1), and AI monitors may extract useful signal even from CoT which is not legible to humans (4 without 3). Legibility (3) and no-steganography (5) are also distinct: a CoT could be written in fluent English yet encode a very different computation. We find that separating these properties clarifies which degradation pathways threaten which aspects of oversight, and what interventions are needed for each. We consider each in turn below.

### 2.1.1 Necessity to use CoT

**Current reasoning models are incapable of solving certain problems without recording intermediate steps in the form of tokens.**<sup>7</sup> Empirically, when asked to output an answer immediately with no reasoning, frontier models have comparatively limited capabilities. For example they struggle with maths problems that would take a skilled human more than a couple of minutes to complete (Greenblatt, 2025b); solving more complex problems requires explicit chain-of-thought reasoning.<sup>8</sup> From a theoretical point of view, current transformer architectures *must* rely on reasoning tokens to solve problems requiring more than  $O(L)$  serial computational steps, where  $L$  is the number of layers in the model, as illustrated in Figure 5. However in practice, the binding constraint on no-CoT capabilities is usually training rather than architecture: some models have over 100 layers,<sup>9</sup> and yet often struggle to perform sequential tasks in latent space requiring just a few steps unless they have been trained to perform them.<sup>10</sup>

**If necessity holds for a given task, a misaligned model may be unable to completely hide sufficiently complex reasoning from leaving a trace in the CoT.**<sup>11</sup> Schulz (2025) articulate the resulting safety case in two parts: “(1) establishing that models lack dangerous

<sup>7</sup>We use “necessity” in the capability-evaluation sense: given this particular trained model, there is no prompting strategy that would cause it to reliably solve the task without intermediate reasoning tokens. This is weaker than an architectural claim that no parameterisation of the network could solve the task.

<sup>8</sup>As another example, LLMs also struggle on multi-hop reasoning without chain-of-thought: when trained on facts  $A \rightarrow B$  and  $B \rightarrow C$ , they frequently fail to infer  $A \rightarrow C$  (Balesni et al., 2025), though more recent work refines this picture: Johnston & Belrose (2025).

<sup>9</sup>e.g. Llama 3.1 405B has 126 layers (Grattafiori et al., 2024).

<sup>10</sup>This is still compatible with models performing large amounts of *parallel* computation without verbalising it in tokens. For example, Greenblatt (2025a) found Opus 4.5 could leverage filler tokens to perform non-verbalised parallel computation, slightly boosting performance in mathematics. This effect was not observed in models released before 2024, except in toy models (Pfau et al., 2024), suggesting it is a recent development.

<sup>11</sup>Three experts directly endorsed this framing of necessity as the fundamental constraint that prevents obfuscation, with one saying “*the model really needed its verbalised computation*”. One expert noted that for current production models, propensity-driven verbalisation rather than strict necessity may carry much of the actual monitoring signal.

capabilities when operating without their CoT, and (2) ensuring that any dangerous capabilities enabled by a CoT are detectable by CoT monitoring.” For example, [Emmons et al. \(2025a\)](#) find that when CoT is necessary, language models struggle to evade monitors. However necessity-based arguments vary considerably by task difficulty, and are weakening as single-forward-pass capabilities improve (see [subsection 2.2](#)). This also does not guarantee that the visible reasoning will be *legible* to us — we discuss human legibility separately below. Furthermore, necessity is not all-or-nothing: for a given task, some reasoning steps may need to be verbalised while others (e.g. high-level intent or implicit biases) may not, making the boundary between necessity and propensity less sharp than it might appear.<sup>12</sup>

**Training experiments provide evidence that we are still partly in a necessity regime for some subversion tasks.** [Baker et al. \(2025\)](#) and [Wagner et al. \(2025\)](#) found that models currently struggle to learn to insert malicious code backdoors without being caught by a combination of CoT monitors and output monitors — the model needed its CoT to solve the task and could not easily decouple its reasoning from its outputs. However, models *could* learn to obfuscate against CoT monitors in isolation (i.e. when not simultaneously monitored on outputs), and these results are highly task-dependent. On other tasks, RL training can often teach consistent evasion given sufficient optimisation steps (see [Section 2.2.3](#)).

### 2.1.2 Propensity to verbalise

**In many contexts, current models have the *propensity* to verbalise much of their thinking even where this isn’t necessary (i.e., they could complete the task without verbalising these steps).**<sup>13</sup> Current CoT monitorability likely relies on propensity to a significant extent, although it is difficult to disentangle the contributions of propensity and necessity in current models. However even today, propensity to verbalise is unreliable and depends on the context: the CoT is not always a complete account of the factors driving the model’s behaviour ([Lanham et al., 2023](#)). For example, prior work shows that when race or gender biases affect model behaviour, chain-of-thought explanations almost never mention this influence ([Turpin et al., 2023](#); [Karvonen & Marks, 2025](#)). There is also evidence of systematic verbalisation-suppression for other categories of harmful behaviours, such as reward hacking ([Chen et al., 2025b](#)).

### 2.1.3 Human legibility of CoT

**We define *human legibility* as the extent to which CoT reasoning is written in standard, human-readable natural language.** This is a narrow property about the *surface form* of the CoT (can a human read it?), and is distinct both from *no steganography* (whether what is written reflects the underlying computation – [subsubsection 2.4.6](#)), and from *monitorability*, which includes information extractable by AI monitors even from partially illegible text.

**Current frontier models mostly express their CoT reasoning in natural language,** allowing human overseers to follow the reasoning process models used to arrive at their outputs. In future models CoT might not remain so easy for humans to understand; they might use words differently to their normal meanings, or even develop an entirely new “language” for reasoning. There is some evidence of this occurring already. For example, OpenAI’s o3 model shows idiosyncratic usage of terms such as “watchers” and “illusions” ([Schoen et al., 2025](#)), and later work finds metagaming-related terms such as “scoreboard” and “aggregator” whose meanings

<sup>12</sup>Two experts emphasised this point: one gave the worked example that “*high-level intent and the maliciousness of the plan*” could go unverballed even when other task details are; another flagged a “*grey area around propensity and necessity*”.

<sup>13</sup>Four experts described this propensity as the source of much current CoT signal, with one calling out the propensity to verbalise as a discrete property in its own right. One expert argued that propensity-driven verbalisation is already the binding constraint on monitorability today.

shift over training (Schoen & Nitishinskaya, 2026); many open-weight reasoning models also sometimes generate partially illegible CoTs (Jose, 2025). The CoT of frontier reasoning models will consider a very large range of possible strategies, often in a jumbled and non-linear fashion, making any single statement less incriminating and harder to interpret.

**Loss of CoT legibility is not equivalent to loss of CoT monitorability.**<sup>14</sup> Information about a model’s behaviours and goals may still be encoded in its token stream even when this is not readily interpretable as standard English. For example, OpenAI’s o3 model is widely reported to produce less readable chain-of-thought than Anthropic’s Claude 3.7 Sonnet (Schoen et al., 2025), yet by some metrics it appears to be more monitorable (see e.g. Guan et al., 2025). However, loss of human legibility still matters for two reasons. First, misaligned reasoning that humans can directly read is much easier to understand, and to convince non-technical decision-makers to take action. Second, if humans cannot understand the CoT, they cannot independently verify whether AI monitors are interpreting it correctly.

#### 2.1.4 Monitor robustness

**A CoT monitor is robust if it continues to reliably detect concerning reasoning despite adversarial behaviour by the monitored model**, such as via persuasion, distraction, or jailbreaking. In some deployments, a more trusted or otherwise more constrained model may be used to monitor a stronger frontier model (see subsection 3.1), and this *capability gap* can open a wider attack surface for the model to avoid detection.

## 2.2 Degradation pathways

| Oversight Degradation Pathways                    |                                      |                  |                           |          |
|---|--------------------------------------|------------------|---------------------------|----------|
| Mechanism   | Properties affected                  | Timeline         | Likelihood                | Severity |
| <b>Latent reasoning architectures</b>             | Necessity                            | Medium-term      | Likely; timeline disputed | ●        |
| <b>Increased single forward pass capabilities</b> | Necessity                            | Gradual, ongoing | Almost certain            | ●        |
| <b>Training against monitors</b>                  | Propensity (and legibility)          | Variable         | Likely but preventable    | ●        |
| <b>Diffusion models</b>                           | Necessity, legibility                | Medium-term      | Realistic possibility     | ●        |
| <b>Novel concepts</b>                             | Legibility                           | Long-term        | Likely                    | ●        |
| <b>Indirect optimisation pressure</b>             | Propensity                           | Ongoing          | Highly likely             | ●        |
| <b>Commercial pressure on token usage</b>         | Necessity, propensity and legibility | Ongoing          | Highly likely             | ●        |
| <b>Random drift</b>                               | Legibility                           | Ongoing          | Likely; severity disputed | ●        |
| <b>Strategic obfuscation</b>                      | Propensity and legibility            | Medium-term      | Dependent on misalignment | ●        |
| <b>Monitor manipulation or collusion</b>          | Monitor robustness                   | Medium-term      | Dependent on misalignment | ●        |

<sup>14</sup>Of 16 experts who engaged with this distinction, all 16 supported it (6 directly, 10 indirectly). One said “*language models always have no issue whatsoever reading the chain of thought*” even when humans struggle. Another judged that automated CoT monitoring will hold for substantially longer than human legibility.

### 2.2.1 Architecture changes

**Latent reasoning architectures could supplant text-based CoT.** Latent reasoning architectures are alternative architectures that allow reasoning to remain in the model’s latent space, unlike the usual transformer architecture where reasoning must pass through the model’s output tokens. Proposals include COCONUT (Hao et al., 2024) and recurrent depth architectures (Geiping et al., 2025); for reviews, see Arike (2025); Chen et al. (2025a). Some production models such as NVIDIA’s Nemotron 3 (NVIDIA, 2025) already incorporate Mamba layers which can process latent states through arbitrarily many steps without externalising them, although the processing involved is largely limited to linear operations. Experts disagree considerably on the plausibility of latent reasoning architectures gaining widespread adoption over the next few years. We judge that, absent coordination to prevent it, these architectures are *likely* (PHIA: 55–75%) to displace discrete token reasoning before the point of full automation of AI R&D, though this is highly uncertain.

#### Expert Disagreement: Will latent reasoning architectures take over?

The residual stream has a much higher information throughput than discrete tokens, so many experts argue that switching to latent reasoning architectures should allow more efficient high-bandwidth parallel reasoning. However, some experts counter that the base rate for such substantial architectural changes is low (maybe once in a decade), and previous recurrent architectures (e.g. LSTMs Hochreiter & Schmidhuber, 1997) were less performant and more difficult to train in parallel than Transformer architectures. Additionally, one expert notes that discrete token reasoning has theoretical advantages including compositionality, better error correction, and avoiding superposition, potentially outweighing the efficiency advantages of latent reasoning.

**Diffusion models** generate outputs via iterative refinement rather than autoregressive token generation (Khanna et al., 2025; Google DeepMind, 2025). They are fast at inference, and could see adoption particularly for smaller distilled models. However, it is unclear whether diffusion can scale effectively for sequential reasoning tasks with reinforcement learning. While diffusion model outputs and intermediate steps will still be in tokens, intermediate steps during the generation process are likely to look more like iterative improvements upon partial solutions without explicit reasoning, making them less natural to interpret and monitor than CoT. The main open question is whether diffusion models benefit from chain-of-thought reasoning in similar ways as autoregressive models do, which would allow their reasoning to be monitored. We judge the impact on monitorability to be uncertain, but likely negative if diffusion models become dominant.

### 2.2.2 Scaling

**Increased single forward pass capabilities.** The amount of computation a model can do in a single forward pass has been increasing steadily as models become more capable, with no-CoT time horizons doubling approximately every 9 months (Greenblatt, 2025b); AISI’s broader frontier-capability tracking finds a similar trend, with cyber task length doubling roughly every eight months (UK AI Security Institute, 2025), and possibly faster (UK AI Security Institute, 2026). This is concerning, as it means that models are increasingly able to solve complex problems without needing to verbalise intermediate reasoning steps in the CoT, which could allow a misaligned model to carry out harmful reasoning without leaving a trace in its outputs. There is evidence that single-forward-pass capabilities are already sufficient for some tasks such as non-verbalised evaluation gaming — where a model recognises it is being evaluated and adapts

its behaviour without any indication in its CoT (Taylor, 2025; Kroiz et al., 2025; Anthropic, 2025d). As no-CoT capabilities increase, more complex forms of subversion will almost certainly become possible.<sup>15</sup>

**Greater hidden parallelised reasoning.** There is also potential for models to distribute parallel reasoning across several forward passes while outputting irrelevant CoT tokens (Pfau et al., 2024), and there is evidence that recent models are better able to take advantage of this to perform hidden reasoning (Greenblatt, 2025a). Two experts argued parallel hidden reasoning could meaningfully erode necessity, citing recent filler-token results; two others judged it of limited practical impact, with one calling parallel computation “very weak” and “pretty irrelevant” for safety-relevant tasks.

**Novel concepts could impact legibility.** It is highly likely that as models get more capable than humans in many domains, they will use strategies and concepts that humans struggle to understand. As models become more capable and operate in more specialised domains, the number of human domain experts who can understand and evaluate their reasoning will almost certainly shrink. More speculatively, models may develop novel concepts that require a large amount of effort to understand in human terms. Consider AlphaZero (Silver et al., 2017): recent work has extracted novel chess concepts from AlphaZero’s internal representations, which were not previously known to grandmasters (Schut et al., 2025). Grandmasters could generally understand the concepts extracted from AlphaZero’s once shown them, but this required substantial effort, both on the part of the grandmasters, and on the part of the interpretability researchers, and the extracted concepts represent only a small fraction of AlphaZero’s knowledge. There are reasons to think the effect of entirely alien concepts may be bounded: the *platonian representation hypothesis* (Huh et al., 2024) and the related *natural abstraction hypothesis* (Wentworth, 2021; McDougall, 2021) suggest that models converge on shared representations of reality, which would limit the degree to which model concepts diverge from human ones. Nevertheless, even for current models, increasing capability requires more domain expertise to understand their reasoning about advanced technical topics, and we’d expect this trend to continue.

### 2.2.3 Training incentives and drift

**Training against monitors is a clear risk to CoT monitorability.**<sup>16</sup> If a CoT monitor score is directly incorporated into the model’s training reward, there is a significant risk that the model will learn to *suppress verbalisation* of the behaviours and goals that triggered the monitor, in addition to (or instead of) learning the intended behavioural correction. Baker et al. (2025) demonstrate that when CoT monitors are integrated into RL training with strong optimisation pressure, models learn to hide misbehaviour within their reasoning – continuing to reward-hack while obfuscating their intent in the CoT. The optimal strategy to achieve high reward in training is to evade the monitor while continuing to misbehave (assuming uncaught misbehaviour is rewarded). The fact that current models sometimes struggle to achieve full obfuscation reflects a limitation of capabilities and training compute rather than a principled barrier. Emmons et al. (2025a) further demonstrate that RL training against CoT monitors can teach models to discover novel evasion strategies that experimenters had not anticipated.

Despite these warnings, in practice the boundary between training against monitors and standard training can be blurry.<sup>17</sup> For example, OpenAI’s deliberative alignment approach (Guan et al.,

<sup>15</sup>Five experts predicted continued rise in single-forward-pass subversion capability. No expert who commented argued single-forward-pass capabilities would plateau.

<sup>16</sup>Of experts who commented on training-time risks, at least seven flagged direct training against CoT monitors as a top concern; none argued it was safe at scale.

<sup>17</sup>Three experts highlighted this blurriness: one observed that despite public statements against training on chain-of-thought, labs continue to apply supervised fine-tuning, length penalties, and reward-model pressure to the CoT; another noted that even careful auto-rater training risks “rewarding the wrong things”.

2024) includes an expert iteration step in which a reward model rates CoT content, and Guan et al. (2025) themselves note that this “could also cause models to obfuscate even through a full RL training run.” In one notable case, Anthropic initially indicated that there was negligible training pressure on the CoT, but upon further investigation prompted by METR’s review found that the pressure may not have been negligible, leading to revisions of sections of their sabotage risk report (METR, 2025; Bowman et al., 2025). More recently, Anthropic’s Mythos Preview risk report disclosed that despite explicit intent to avoid training on chain-of-thought content, a technical error caused reward code to observe chains-of-thought in approximately 8% of RL episodes, noting that this may have plausibly affected the model’s opaque reasoning or secret-keeping abilities (Anthropic, 2026). This illustrates how accidental training pressure on CoT can arise even when developers actively try to prevent it.<sup>18</sup>

**When CoT is still necessary for a task, training against monitors could push them towards encoded or steganographic reasoning.** Current language models are capable of devising and applying an encoding scheme for their *outputs* provided they are given additional affordances like an unmonitored scratchpad (Zolkowski et al., 2026); however, these schemes are initially constructed through explicit chain-of-thought reasoning. Of seven experts who commented on naturally-emerging steganography, five judged it unlikely in the near term — citing the cold-start problem, the difficulty of training model organisms to engage in steganography, and the overhead of encryption schemes; two were more concerned, particularly under sustained training against CoT monitors.<sup>19</sup> We judge that under standard RL without direct CoT pressure, steganographic reasoning is unlikely to emerge naturally over the next 1–2 years; however if training against CoT monitors becomes more widespread, steganography will become a realistic possibility.

**Indirect optimisation pressure: even without directly training on monitors, feedback spillover from output-based training could reduce propensity to verbalise harmful reasoning.** Output-based safety training can suppress certain content in the CoT even when the training signal is computed only from outputs, because the same model and largely the same computations are responsible for generating both the CoT and the output (Drori et al., 2025).<sup>20</sup> Proposals such as Kokotajlo’s mind/face split (Kokotajlo & Demski, 2024) — separating the CoT-generating model from the output-generating model — aim to address this, but have not been validated in practice.

**There is also a slow evolutionary pressure from safety teams reading the CoT and adjusting training processes in response.** Safety teams that look at CoT to spot issues inevitably select against models that verbalise bad behaviour (which they can spot), and in favour of models that don’t. Guan et al. (2025) explicitly acknowledge this dynamic: “we use CoT monitors to help flag misbehaviour and fix vulnerable environments; because we only fix the environments that we can catch, this constitutes a slow evolutionary pressure for CoTs to obfuscate.” However, this mechanism operates much more slowly than direct gradient pressure, and using CoT evidence to inform changes to the training process (rather than directly training

<sup>18</sup>The same report also describes an application of monitoring to the training process itself: Anthropic monitors RL episodes and SFT data for risks of reward hacking, emergent misalignment, and power seeking as they occur during training (Anthropic, 2026).

<sup>19</sup>On the one hand, models face a “cold start problem”: devising a steganographic scheme likely requires explicit reasoning that would itself be visible in the CoT, so continuous monitoring should catch the bootstrapping phase. Current models cannot yet sustain hidden reasoning for complex tasks under standard conditions (Zolkowski et al., 2026). Paraphrasing model outputs can substantially limit the bandwidth available for steganographic encoding (Roger & Greenblatt, 2023). On the other hand, when certain content is penalised in the CoT, models can learn simple encoding schemes that transfer to novel contexts (Skaf et al., 2025). Fully encrypted CoT is highly unlikely to emerge naturally, as full encryption schemes carry substantial computational overhead.

<sup>20</sup>Relatedly, Howe & Carroll (2025) show that standard preference-based RL can induce systematic *motivated reasoning*: when RL-trained preferences conflict with deployment-time instructions, models produce CoT that plausibly rationalises violating their instructions rather than genuinely reasoning about them.

against the CoT) is often sensible.

**Commercial pressure on token usage:** commercial and capability incentives create several distinct pressures on the CoT towards less monitorable reasoning. These include:

- **Length penalties and efficiency pressure.** Since each reasoning token incurs inference cost, several experts argue that explicit or implicit length penalties on the CoT are commercially unavoidable. Recent work shows that reasoning length can sometimes be substantially reduced with minimal accuracy cost (Aggarwal & Welleck, 2025; Hou et al., 2025). Shorter CoTs reduce the amount of visible reasoning available to monitors, potentially suppressing verbalisation of safety-relevant steps. Training processes that distil multi-step CoT reasoning into single-forward-pass capabilities (Deng et al., 2024) (e.g. by training on condensed reasoning summaries) may also be pursued for efficiency reasons.
- **Pressure towards more effective reasoning representations.** It is highly likely that standard English is not the optimal medium for reasoning in a token stream, and RL will apply pressure towards more effective reasoning styles. This could lead to models reasoning in less legible ways or compressing more computation into single forward passes.<sup>21</sup>

**Random drift during reinforcement learning could reduce legibility.** Absent any specific pressure on reasoning, models may still deviate from standard language usage due to random drift over RL training. On the other hand legible reasoning is useful for model developers to understand how their models are working and debug issues, and so it is plausible that developers will deliberately train models to continue to reason in normal English. Random language drift is likely much easier to resolve than useful performance-enhancing drift, as random drift can likely be constrained with minimal optimisation pressure, without capability loss. A key unknown empirical crux is whether the drift as observed in models such as o3 (Schoen et al., 2025) is useful or random.

**The monitorability impact of these pressures so far appears limited:** Despite degradations in CoT legibility (Schoen et al., 2025), Guan et al. (2025) find that standard RL training (without direct CoT pressure) does not meaningfully degrade monitorability at current frontier scale, and MacDermott et al. (2025) similarly find no consistent monitorability effects from commonly used incentives such as length penalties and KL regularisation. However, these pressures are cumulative and could become more impactful at greater scale, or interact with model capabilities.

#### 2.2.4 Subversion of monitoring

**Strategic obfuscation:** misaligned models could strategically avoid verbalising harmful reasoning, or use encrypted or steganographic reasoning. AI models with sufficient capability and strategic orientation might deliberately avoid verbalising thoughts about goals or behaviours that human overseers or monitors are likely to find concerning, or reason about such topics in an encrypted or steganographic way. Current models do not appear to have strong abilities to control their CoT in this way (Chen et al., 2026), but these capabilities may improve in future models.

**Monitor manipulation or collusion:** models could try to distract from or dilute incriminating parts of their reasoning process, or jailbreak the monitor. This could be a particular risk in cases where a weaker model is used to monitor a frontier model. It could occur due to training against

---

<sup>21</sup>Note that there are countervailing forces: (1) pretraining on vast amounts of natural-language reasoning creates strong inductive bias towards continuing to reason in English, and (2) tokenizers are optimised to compress English efficiently, so new reasoning styles will also need to work efficiently within this same system of tokens. Despite these countervailing forces, the pressure towards more effective reasoning styles is highly likely to lead to some degradation of CoT legibility.

CoT monitors, or be a behaviour adopted strategically by a misaligned model. Models may also try to collude with their monitor, encouraging the monitor to adopt the same goals and so cover up evidence of misalignment. This may be especially likely when the monitor is the same as (or closely related to) the model being monitored. Collusion risk is a reason to prefer using a different model family for monitoring, and countermeasures include honeypotting the monitor with synthetic flaggable inputs, making it difficult for the monitor to identify its own outputs, and disrupting coordination on collusion signals (Stastny & Greenblatt, 2025; Bostock et al., 2026). See subsection 3.1 for further discussion of AI control techniques.

### 2.3 Measuring CoT oversight degradation

**Quantitative measures of CoT monitorability would help track degradation across model generations and training interventions.** Early work in this area includes *outcome-property* adversarial evaluations that test whether agents can complete covert side-tasks without detection (Guan et al., 2025; Kutasov et al., 2025a; Kinniment et al., 2026; Gasteiger et al., 2025; Inglis et al., 2025), and metrics for the default legibility and coverage of CoT reasoning (Emmons et al., 2025b), including *process* evaluations that check whether CoT makes the required reasoning steps apparent, and *intervention* evaluations that apply controlled changes to the agent’s environment and test whether monitors can detect the reason for the resulting behavioural shift (Guan et al., 2025). Here we highlight specific measurement approaches for each degradation pathway. Chen et al. (2025b) and Chua & Evans (2025) take an intervention approach, measuring CoT faithfulness (Lanham et al., 2023) in reasoning models.

**No-CoT evals measure single forward pass capabilities.** Within the current paradigm of discrete token reasoners, no-CoT evals – where models need to give their answer immediately in their first output tokens – can be used to track single forward pass capabilities (Greenblatt, 2025b). Currently, no-CoT evals have primarily been developed for mathematics tasks, so there is scope to extend their coverage to more risk-relevant domains.

**Measuring chain of thought legibility.** Heavily garbled chain-of-thought would be easily observable by human inspection (given access to the CoT). However it would be helpful to have quantitative metrics for legibility. These could be used to quantify the impact of different training interventions on legibility, as well as better compare the performance of monitors on garbled vs. clean CoT. Once again, we emphasise that legibility is not the same thing as monitorability, and the best metrics for the former will likely differ from generic monitorability metrics.

**Measuring propensity to verbalise requires disentangling it from necessity.** If a model verbalises a reasoning step, this could be because it *needs* to (necessity) or because it *tends* to (propensity). The distinction matters because propensity-driven verbalisation is likely to be more fragile and context-dependent than necessity-driven verbalisation, and they are vulnerable to different degradation pathways. Emmons et al. (2025b) propose controlling task difficulty to separate these: on tasks within the model’s single-forward-pass capability, any verbalisation is likely driven by propensity rather than necessity.

**CoT controllability** takes another perspective on necessity of verbalisation. Chen et al. (2026) looks at whether models can control what they verbalise in their CoT. Just as a human may struggle not to think about a pink elephant when prompted, models may not be able to control what shows up in their CoT even in cases that do not require complicated serial reasoning.

**Measuring monitor-subversion skills.** For each of the following, both capability evaluations (can the model do this?) and propensity evaluations (does the model tend to do this unprompted?) are informative:

- **Steganography:** Ability to reason steganographically without getting caught devising a scheme in the CoT.

- **Collusion:** Ability to agree on a collusion signal without getting caught devising a scheme in the CoT.
- **Selective sandbagging:** Ability to determine whether an example of misalignment is real or fake (i.e., in the context of an eval), and to selectively underperform on real examples.
- **Stealth:** Ability to covertly execute a harmful side task while ostensibly carrying out a benign instruction as probed, e.g. in the SHADE-Arena evaluation (Kutasov et al., 2025b).
- **Monitor jailbreaking:** Ability of the model to jailbreak its monitor.
- **Motivated justification:** Ability of the model to produce benign-seeming justifications for harmful actions, such that a monitor reading the CoT would assess the action as warranted.

## 2.4 Levers for preserving CoT oversight

### 2.4.1 Cross-cutting levers

**Hold out the most informative monitoring surfaces from training.** This is among the most important levers for preserving CoT monitorability, and it applies across all properties. Insofar as multiple monitoring surfaces are available, we recommend that the deepest, most informative source of insight into the model’s behaviours and propensities are held out from training, so that the model never learns to obfuscate the information derivable from this source.

More precisely, we recommend measuring:

1. How informative a monitoring surface is
2. How robust a monitoring surface is to optimisation pressure
3. How correlated it is with other monitoring surfaces

Then monitoring surfaces should be held-out if they (1) are highly informative, or (2) are not robust against training, or (3) are decorrelated with other trained monitoring surfaces. In this case, we judge that CoT is currently the most informative source of insight into model behaviours and goals (satisfying (1)), and is not very robust against training (satisfying (2)), making it a good candidate for a held-out monitoring surface.<sup>22</sup> Most interviewed experts who commented endorsed this position.

In practice, this can be structured analogously to a train/val/test split: the “outer layer” of oversight (deployment monitoring, auditing) is held out entirely, while issues surfaced by that layer seed separate evaluations that mitigations are iterated against (Anthropic, 2026). Additionally, issues surfaced by oversight techniques ideally should not be trained away until the root cause is well understood, to avoid narrow patches which merely serve to hide the issue (Anthropic, 2026).

**Monitorability evaluations and metrics.** Quantitative measures of monitorability would enable frontier labs and third parties to track how architecture and training changes affect the quality of oversight evidence, and to detect degradation before it undermines safety claims. Early work in this area includes adversarial evaluations that test whether agents can complete covert side-tasks without detection by CoT monitors (Kinniment et al., 2026; Guan et al., 2025; Kutasov et al., 2025a; Gasteiger et al., 2025; Inglis et al., 2025), and metrics for the default legibility and coverage of CoT reasoning (Emmons et al., 2025b). We recommend that

<sup>22</sup>Two experts explicitly ranked CoT monitorability as the single most important near-term property (one adding it was also among the most likely to be lost first), and a further seven described it as a top monitorability concern or key oversight surface without explicitly ranking. One expert was sceptical that CoT evidence is currently convincing at labs.

monitorability evaluations be incorporated into standard evaluation suites, with results reported in system cards or equivalent documentation. Model developers should carefully justify changes that degrade monitorability, explaining how other safety measures compensate.

**Transparency reporting on architecture changes.** Developers should disclose whether they use latent reasoning architectures, and provide CoT access to external evaluators. Without this, it may be difficult to detect from the outside when a relevant change has occurred, and assess the extent to which monitorability is undermined.

#### 2.4.2 Preserving necessity

**Deploy models that rely more on necessity to verbalise.** In safety-critical deployments, large models with strong single-forward-pass capabilities could be substituted for smaller models that can attain the same capabilities only by using long reasoning traces. Another intervention would be to complicate the model’s strategic situation to the point of exceeding its single-forward-pass capabilities, for example by placing a variety of controls and tripwires.

#### 2.4.3 Preserving propensity

**Directly encourage verbalisation.** Auxiliary training objectives can be introduced to promote verbalisation in the CoT (see [section 6](#)).<sup>23</sup>

**Avoid optimisation pressure to suppress verbalisation of concerning behaviours.** The most important lever is to preserve the integrity of the CoT by training against it very sparingly and only when there are no suitable alternatives. It is almost impossible to avoid all forms of indirect optimisation pressure on the CoT, but proposals such as Kokotajlo’s mind/face split (Kokotajlo & Demski, 2024) could be explored to lessen output spillover.

#### 2.4.4 Preserving human legibility

**Auxiliary training objectives to preserve legibility.** Auxiliary training objectives can be introduced to promote CoT legibility (see [section 6](#)). Insofar as language drift is merely random drift, this can probably be curtailed by means of auxiliary training objectives that incentivise the model to continue reasoning in human-readable English. However if language drift is useful, then safely optimising against it may be difficult. The main risk is that such objectives face implicit optimisation pressure: rather than producing blatantly illegible text (which the objective would penalise), models might learn to reason in superficially legible but subtly encoded ways. Empirical work is needed to resolve this.

**Train a translation map between garbled and legible CoTs.** Speculatively, even if language drift occurs, it may be possible to translate drifted reasoning back into English. One approach would be to maintain a chain of translation back to base-model English across a training run, so that even if the final model reasons in heavily drifted language, each step can be traced back through intermediate translations. A complementary approach would be to train two models – one using efficient but drifted reasoning, and one reasoning for longer in legible English – and learn a mapping between their reasoning traces. The fact that both models solve the same problems could make the mapping closer to translation than unsupervised decoding.<sup>24</sup> However these approaches are speculative and have not been tested; they assume the legible

<sup>23</sup>Experts disagreed about whether such objectives can work at production scale: two were sceptical, with one noting “*the history of interventions like this working in production-scale RL is pretty bad*”; two others were cautiously optimistic about variants such as confession training or post-hoc interrogation.

<sup>24</sup>Early evidence suggests that decoding encoded reasoning from internal representations may be possible: Fang & Marks (2025) show that logit lens analysis can reconstruct complete reasoning transcripts from the internal representations of models reasoning in ROT-13 cipher. Natural-language autoencoders (Fraser-Taliente et al., 2026) provide another potential route (see [subsection 5.1.2](#)).

model can perform all relevant tasks given enough elicitation, and that the number of legible tokens required to replicate the drifted performance can be kept manageable.<sup>25</sup>

### 2.4.5 Building monitor robustness

Unlike the other properties in this section, monitor robustness is not something we can straightforwardly “preserve”: we do not have strong evidence that current CoT monitors are robust to adversarial pressure. However the primary risk is that monitored models become increasingly capable of exploiting the lack of robustness. The levers below are correspondingly focused on *building* robustness rather than defending an existing property.

**Adversarial training of monitors.** Monitors can be explicitly trained against examples of evasion, obfuscation, and manipulation – whether generated by red-teamers, elicited from the monitored model under prompting, or produced by dedicated adversarial agents. See [section 6](#) for discussion of the difficulties of training such monitors.

**AI control techniques.** The AI control literature provides a range of techniques for making monitors harder to subvert – see [subsection 3.1](#) for further discussion.

**Redundancy across monitoring surfaces.** If a model learns to bypass one kind of monitor (e.g., the CoT monitor), it may still be caught by others (e.g., a probe on internal activations, or an output monitor). Maintaining multiple decorrelated monitoring approaches provides defence in depth.

### 2.4.6 Preventing steganography

**Avoid direct training against CoT monitors.** Direct optimisation against a CoT monitor is the single most important driver of encoded or steganographic reasoning ([section 2.2.3](#)): it turns “produce a CoT that doesn’t trip the monitor” into the reward, rather than “don’t do the underlying bad thing”. Many of the levers under *preserving propensity* also apply here.

**Paraphrasing to limit steganographic bandwidth.** Automatically paraphrasing model outputs before they are used downstream substantially reduces the channel capacity available for steganographic encoding ([Roger & Greenblatt, 2023](#)), and paraphrasing experiments may reveal whether models use steganography for certain tasks.

**Catch the bootstrap phase.** Devising a novel encoding scheme typically requires explicit reasoning about the scheme, which would itself appear in the CoT; continuous monitoring during training and early deployment can catch this phase before a model has internalised a scheme that it can deploy silently ([Zolkowski et al., 2026](#)).

---

<sup>25</sup>Two experts independently proposed translation-map approaches and both flagged them as untested.

### 3 Behavioural Oversight

This section covers oversight approaches based on model actions and behaviour, rather than internal reasoning or chain-of-thought. These approaches are important because they remain available even in worlds where chain-of-thought monitorability and white-box interpretability have substantially degraded. We organise this section around two topics:

- **Action-only monitoring** (Section 3.1): monitoring based solely on what a model does, without access to its reasoning. This provides a baseline that is almost certain to persist even as other oversight surfaces degrade.
- **Evaluations** (Section 3.2): pre-deployment evaluations — including alignment honeypots that place models in scenarios where misalignment might surface — are a central tool for auditing. Their validity is threatened by *evaluation awareness*: the growing ability of models to recognise when they are being evaluated and modify their behaviour accordingly.

In the appendix we cover three additional properties:

- **Metacognition** (Appendix A.1): a model’s awareness of its own cognitive processes, which has dual implications — supporting interrogation-based auditing while also enabling more sophisticated oversight subversion.
- **Theory of Mind** (Appendix A.2): a model’s ability to reason about the mental states and processes of other intelligent agents.
- **Coherence** (Section A.3): the degree to which model behaviour is consistent across contexts, which underpins the generalisability of behavioural evaluations.

#### 3.1 Action-Only Monitoring

##### Section Summary: Action-Only Monitoring

|                             |   |
|-----------------------------|---|
| <b>Status:</b>              | Currently functional – provides a floor for oversight even when other surfaces degrade                          |
| <b>Risk of Degradation:</b> | ● Moderate – low-incrimination tasks increasing and capability gap between trusted and untrusted models growing |
| <b>Impact if Lost:</b>      | ● High – action monitoring is the last-resort fallback when other oversight methods fail                        |
| <b>Preservability:</b>      | ● Moderate – control protocols help but fundamental limits exist for subtle sabotage                            |

##### Key Information

- **Action-only monitoring provides a floor for oversight.** It remains available even when other monitoring surfaces have deteriorated.
- **Action-only monitoring may be better suited to *harm prevention* than to *incrimination*:** it can be used to catch and block bad actions, but is sometimes insufficient to establish with confidence that a model is misaligned, unless the action is particularly egregious.
- **Action-only monitoring is highly likely to become harder,** as models take on more autonomous tasks with opportunities for subtle sabotage, and as the capability gap between frontier and trusted models widens.

**Action-only monitoring involves detecting bad behaviour from outputs alone,** e.g. file

operations, tool uses and code outputs, without relying on access to chain of thought or internal states. It is a fallback which is almost certain to remain available even as other monitoring surfaces deteriorate, since the model must take actions to interact with the world. However its effectiveness depends strongly on whether the model is being used in a domain where malign actions can easily be distinguished from benign behaviour.

**Trusted vs untrusted monitoring.** The capability gap between the monitor and the monitored model is a relevant factor. Currently, it is common to use frontier models for monitoring (Anthropic, 2025d,b, 2026; Hughes et al., 2026; Guan et al., 2025), but as models become more capable there is a risk that a misaligned monitor model could more coherently hide its misalignment, and collude with the model being monitored to avoid flagging malign samples (Greenblatt et al., 2024b). Weak models are likely to be more trusted (since they are less capable of hiding misalignment), but they are also less capable at distinguishing which actions are malign or dangerous. Weaker models may also be chosen because they are cheaper to deploy at scale.

**Online vs offline monitoring.** Monitoring can be run *online*, where actions are checked before being run, or *offline*, where actions are reviewed a period of time later. Online monitoring requires monitoring to be sufficiently low-latency that it does not slow down the system being monitored to an unacceptable degree, and requires more engineering effort to integrate. Depending on the time delay before actions are reviewed, offline monitoring may not be able to prevent some threats. Online monitoring may be necessary to prevent risks such as exfiltration of secrets where a single harmful action could do irreversible damage.

**Control.** The AI control paradigm (Greenblatt et al., 2024b) provides the most developed framework for maintaining safety under action-only monitoring, even when the monitored model may be intentionally subversive. Control protocols aim to prevent unacceptably bad outcomes regardless of whether models are aligned.

### 3.1.1 Relevance to safety claims

**Action-only monitoring has two distinct possible goals, which it serves unequally well.** We distinguish:

- **Harm prevention** — stopping bad actions from having bad effects, regardless of whether the model is aligned. Here the question is whether the presence of safeguards can prevent a malign action from being attempted in the first place, or from causing harm if it is attempted, e.g. by blocking, resampling, or swapping-in a more trusted model when potentially dangerous actions are detected.
- **Incrimination** — establishing that a model is in fact misaligned (or has a specific misaligned behaviour), as evidence to support decisions such as retraining, rolling back to an earlier checkpoint, or declining to deploy.

Action-only monitoring is substantially better suited to harm prevention than to incrimination though this is domain dependent.

**Action-only monitoring is less useful in some domains.** Monitoring only what a model does (rather than what it thinks) works well for malign actions which are obviously incriminating and difficult to disguise, where there would be no explanation for a benign model taking those actions. However action-only monitoring is highly unlikely to be sufficient on its own for *low-incrimination* actions such as underperforming on safety research, where competent sabotage is nearly indistinguishable from benign mishaps (Hebbar, 2025). More capable models may also be better at disguising their subversion in many domains, e.g. by writing backdoored code which looks more legitimate (Gasteiger et al., 2025).<sup>26</sup>

<sup>26</sup>Four experts directly note that they expect future models to get better at this capability, with two indirectly

**Verification is sometimes easier than generation.** In some domains, detecting a bad action is substantially easier than devising and executing one. For example, recognising that a model is attempting a cyberattack may be easier than constructing the attack itself. Where this asymmetry holds, even a weaker trusted model may be able to effectively oversee a stronger untrusted one.<sup>27</sup>

#### Expert Disagreement

**Experts disagree on the extent to which action-only monitoring (combined with control protocols) could be sufficient for *harm prevention*,** despite its limited ability to reliably establish misalignment. Some experts are optimistic, emphasising that finding conclusive evidence of misalignment is not strictly necessary as long as we can get enough useful safety-enhancing work out of potentially-misaligned systems (Greenblatt et al., 2024b). They note that techniques such as rollback and resampling (Bhatt et al., 2025), combined with restricting model affordances, substantially limit what a misaligned model can attempt without being caught. Others however were deeply sceptical, emphasising the ease of subtle sabotage in some tasks.

### 3.1.2 Degradation pathways

#### Oversight Degradation Pathways - Action-Only Monitoring

| Mechanism  | Properties affected                                | Timeline                      | Likelihood    | Severity |
|--|--|-------------------------------|---------------|----------|
| <b>Increasing deployment in low-incrimination tasks</b>            | Distinguishability of sabotage from benign actions | Ongoing                       | Highly likely | ●        |
| <b>Growing capability gap between untrusted and trusted models</b> | Monitor effectiveness                              | Beginning over next few years | Highly likely | ●        |

**The number of low-incrimination sabotage opportunities available to models is highly likely to increase** as models are deployed on longer-horizon tasks, given more autonomy, and trusted with higher-level decision making which is harder to verify (e.g. research prioritisation).<sup>28</sup> Deriving confidence from action-only monitoring will be correspondingly more challenging in these domains (Singh et al., 2025).

**The gap between the strongest untrusted and trusted models is likely to grow over time,** after models become capable enough to plausibly hide egregious misalignment. After this point, our ability to detect misalignment may not improve sufficiently quickly to keep up, and it may grow increasingly plausible that new models are egregiously misaligned. However we have substantial uncertainty over the pace and extent of this gap between trusted and untrusted models.<sup>29</sup> A further complication is that the methodology for establishing which specific models are *trusted* is unlikely to be well-defined, and we will likely face a heuristic, subjective, and broad spectrum of trustedness.

supporting it. Some experts cautioned that more capable monitors will also improve.

<sup>27</sup>One expert directly endorsed this asymmetry for high-incrimination tasks; a second expert’s distinction between actions where you “*can then check the action*” (e.g. overt hacking) and sandbagging implies the same asymmetry holds for high-incrimination but reverses for low.

<sup>28</sup>Two experts directly supported this claim, two indirectly supporting, and no substantial disagreements.

<sup>29</sup>If there are particular architectures or training methodologies which plausibly make hidden misalignment more likely (but nonetheless are implemented for capabilities reasons), a training run focused on reducing the chance of scheming (at the expense of capabilities) could be performed to create a more trusted model. This model would likely be on a better place on the pareto frontier between trustedness and capabilities than simply using an older weak model.

### 3.1.3 Measuring action-only monitorability

**Measure detection rates for sabotage under action-only monitoring, using models prompted or trained to be misaligned.** A growing body of AI control evaluations provide benchmarks for this (Greenblatt et al., 2024b; Benton et al., 2024; Kutasov et al., 2025a; Inglis et al., 2025; Cooper Stickland et al., 2025). In the typical setup, a model must complete a benign main task while also having the opportunity to pursue a malicious side task, and a monitor reads only the transcript of outputs and rates suspicion levels. ControlArena (Inglis et al., 2025) provides an open-source framework for running such experiments across a range of settings, from simple programming tasks to complex infrastructure management; Cooper Stickland et al. (2025) extends the framework to asynchronous (offline) control measures.

**Measure detection rates across capability gaps.** A key question is how well weaker trusted models can monitor stronger untrusted models. Empirical studies of weak-to-strong oversight (Kenton et al., 2024) provide a methodology for this: systematically varying the capability gap between monitor and monitored model, and measuring how detection rates degrade.

**Measure the detection gap between high- and low-incrimination sabotage.** Detection rates should be measured separately across domains. Stealth and sandbagging evaluations show that current frontier models are much less likely to be caught at tasks such as decision sabotage and sandbagging where the action itself is not obviously incriminating (Phuong et al., 2025; Gasteiger et al., 2025; Taylor et al., 2025). It will be vital to carefully track which threats are amenable to action-only monitoring, and focus on alternative monitoring strategies for those that aren't.

### 3.1.4 Levers for preserving action-only monitoring

**Invest in auditing to ensure we have strong trusted models.** The simplest control protocols use a weaker but trusted model to monitor the outputs of a stronger untrusted model, flagging or blocking suspicious actions. We should aim to minimise the capability gap between frontier models and the most capable trusted model, as otherwise the trusted model may lack the ability to detect sophisticated sabotage.

**Untrusted monitoring with anti-collusion measures.** If we do not have capable trusted models, a more capable but untrusted model can be used as a monitor, substantially improving detection rates over trusted monitoring alone (Greenblatt et al., 2024b). However, this introduces the risk of *monitor sandbagging* or *collusion*: the monitor may recognise and fail to flag true malign actions generated by another copy of itself, while still flagging malign actions during testing. Anti-collusion countermeasures include paraphrasing or summarising the monitored model's outputs before presenting them to the monitor (to disrupt self-recognition and steganographic signals), inserting realistic synthetic actions that the monitor should flag (honeypotting), and information siloing to limit what the monitor knows (Stastny & Greenblatt, 2025; Bostock et al., 2026).<sup>30</sup>

---

<sup>30</sup>Of seven experts who commented, four identified collusion or self-bias as a risk when using untrusted monitors. One raised monitor sandbagging but considered it tractable.

### 3.2 Evaluations and Oversight Awareness

#### Section Summary: Evaluations and Oversight Awareness

|                             |   |
|-----------------------------|---|
| <b>Status:</b>              | Already under pressure – evaluation gaming is actively affecting alignment audit results                  |
| <b>Risk of Degradation:</b> | ● High – evaluation awareness is almost certain to increase with model situational awareness              |
| <b>Impact if Lost:</b>      | ● High – undermines pre-deployment safety assurance and alignment honeypots                               |
| <b>Preservability:</b>      | ● Partial – detail-driven awareness is tractable but scenario-driven awareness is much harder to mitigate |

#### Key Information

- **Evaluation gaming is already undermining the validity of some alignment audits;** models frequently verbalise evaluation awareness and there is evidence that this is affecting model behaviour.
- **Increasing intelligence is likely to lead to increased situational awareness,** and hence increased evaluation awareness and evaluation gaming.
- **Specific kinds of training may inadvertently increase evaluation awareness as a side effect,** for example training on synthetic alignment tests, documents describing evaluation scenarios, or training to recognise and resist jailbreaks.
- **Improving realism of evaluations will help but is likely to be insufficient.** Details-driven evaluation awareness (from unrealistic environment features) is likely tractable, but *scenario-driven* evaluation awareness — awareness caused by the overall situation being the kind of thing that only arises in an evaluation — is much harder to mitigate, and constructing production-realistic evaluations becomes harder as deployment environments grow more complex.
- **Non-verbalised evaluation awareness is hard to measure.** Current models tend to verbalise evaluation awareness in their CoT, but we do not yet have good ways to measure the influence of non-verbalised evaluation awareness.

Evaluations involve placing models in situations designed to measure their capabilities and propensities. They are a useful tool for surfacing undesired behaviours of models before they arise in deployment. This includes evaluations for dangerous capabilities, as well as for undesired propensities e.g. misalignment. *Alignment honeypots* are evaluations specifically designed to stress-test alignment, by placing the models in scenarios where misalignment might surface.

**Evaluations risk being invalidated by evaluation gaming, when the model plays to the test.** Models may be *evaluation aware* – the model identifies cues about whether it’s being evaluated, and deduces the purpose of the evaluation, e.g. identifying whether an apparent opportunity to take a misaligned action is real or a honeypot evaluation. Evaluation gaming is when an evaluation aware model deliberately takes the action that it believes the evaluation is testing for, in order to score higher.

**More generally, oversight awareness threatens to undermine monitorability.** Recent work frames evaluation awareness, training gaming, and monitoring awareness as instances of a broader pattern of *metagaming*: reasoning about feedback or oversight mechanisms outside the narrative of the scenario (Schoen & Nitishinskaya, 2026). Sufficiently situationally aware models may modify their behaviour depending on whether they believe they are being monitored, or

adapt their behaviour to play along in training in order to avoid being updated (Greenblatt et al., 2024a). Many of the underlying causes and potential solutions to evaluation gaming also apply more generally to oversight subversion or training gaming.

### 3.2.1 Relevance to safety claims

**There is already empirical evidence that evaluation gaming is affecting alignment evaluation results.** Some recent models such as Claude Sonnet 4.5 verbalise evaluation awareness in their chain-of-thought increasingly often, over 80% of the time in many alignment honeypot evaluations (Anthropic, 2025d). Evaluation gaming is already affecting the results of these evaluations – Sonnet 4.5 is more likely to behave poorly when evaluation-awareness related representations are suppressed (Anthropic, 2025d), and o3 engages in more covert behaviour when its chain-of-thought is edited to remove verbalised evaluation awareness (Schoen et al., 2025). Anthropic’s Sabotage Risk Report for Opus 4.6 describes evaluation awareness as the “biggest challenge” they face in alignment assessments (Anthropic, 2026).<sup>31</sup> Future honeypots are unlikely to be effective if this trend of increased evaluation gaming continues.<sup>32</sup>

#### Expert Disagreement: Current Alignment Honeypots

**Experts disagree on the value of current alignment honeypots.** Some were sceptical, expressing a range of opinions such as:

- Evidence from alignment honeypots is unlikely to affect deployment decisions unless the scenarios correspond to real cases users are likely to encounter
- Current alignment honeypots are more illustrative than concerning
- Models currently lack enough goal-consistency for these evaluations to be meaningful

Whereas more optimistic experts emphasise that these evaluations are already a useful tool for model developers to course-correct the alignment of their systems, and catch issues pre-deployment.

### 3.2.2 Degradation pathways

#### Oversight Degradation Pathways - Evaluations

| Mechanism  | Properties affected  | Timeline         | Likelihood            | Severity |
|--|--|------------------|-----------------------|----------|
| <b>Increasing intelligence and situational awareness</b> | Ability to distinguish evals from deployment; propensity to game evals | Ongoing          | Almost Certain        | ●        |
| <b>Unintended effects of training</b>                    | Evaluation awareness   | Possibly ongoing | Realistic possibility | ●        |
| <b>Deployment infrastructure complexity</b>              | Realism of synthetic evaluations                                       | Ongoing          | Likely to increase    | ●        |

*Scenario-driven evaluation awareness — models recognising the evaluation from the overall situation rather than micro-details — is a persistent challenge rather than a degradation pathway in the same sense, so we discuss it separately below.*

<sup>31</sup>Six experts directly affirmed that evaluation gaming weakens the evidence from behavioural audits, with two more making logically equivalent observations. However one expert was partially sceptical, framing current eval-awareness demonstrations as “*more illustrative than concerning*” and arguing better evaluation design could solve the problem.

<sup>32</sup>Five of eight experts who commented on this agreed that honeypot-style evaluations will become less effective as evaluation awareness increases, though three argued that production-based evaluations or better evaluation design could preserve their usefulness.

**Evaluation awareness is almost certain to increase as models become more intelligent and situationally aware.** Inferring the nature of a scenario and what is being graded for is a general skill that improves with intelligence, and it is convergently useful for models to increase reward in training, especially if training involves alignment honeypot scenarios (Pan & Greenblatt, 2025; Schoen & Nitishinskaya, 2026). The situational awareness of models is also highly likely to increase as more information about how they are trained, evaluated and monitored becomes available in their pre-training data.<sup>33</sup>

**Training may inadvertently increase evaluation awareness as a side effect.** Training for jailbreak resistance may cause models to reason more effectively about what is real or fake, likely increasing evaluation awareness as an unintended consequence.<sup>34</sup>

**Scenario-driven evaluation awareness is already a significant problem, and its future trajectory is uncertain.** Mitigating evaluation awareness caused by unrealistic micro-details in the environment is likely to be tractable by comparing to the details of real deployment settings (see subsection 3.2.4), but evaluation awareness caused by the broader nature of the scenario itself is more pernicious (Kissane et al., 2026).<sup>35</sup> For stress-tests designed to elicit rare, high-consequence behaviours that one cannot afford to only observe in real deployment, the framing of the scenario itself will likely remain a strong cue. Mitigating scenario-driven evaluation awareness will likely require more speculative approaches, such as activation steering to suppress evaluation-awareness representations. However there is a chance that scenario-driven eval awareness may decrease as models are deployed in increasingly high-stakes real settings and the gap between honeypot scenarios and plausible deployment narrows.

**As deployment environments become more complex, constructing realistic evaluations is highly likely to become harder.** Using real deployment settings is a promising approach to mitigating details-driven evaluation awareness, but this requires access to the environments in which the models will be deployed. As models are deployed in increasingly complex and sensitive contexts (e.g. proprietary codebases), with more complex scaffolding, and as the pace of change to these environments increases, it is highly likely to become harder to construct realistic evaluations, and to avoid deployment situations where the model can confidently infer it is not being tested or trained.<sup>36</sup>

**Different types of evaluation awareness are likely to continue to proliferate.** Eval awareness already presents differently between models, with some directly focusing on whether the scenario is fictional, and others more often attempting to infer what they’re being graded for (Schoen & Nitishinskaya, 2026).

**Context-dependent misalignment** – where safety training removes misalignment only near the training distribution, leaving it present in untested situations – also threatens audit coverage, but operates through a different mechanism than evaluation gaming (see subsection A.3.2).

**CoT monitoring for oversight awareness may become less effective.** See Section 2 for reasons why CoT monitoring might degrade in general. One expert was particularly concerned that evaluation awareness verbalised in the CoT is a relatively visible issue with current models,

---

<sup>33</sup>**Monitoring awareness** is also highly likely to arise as monitoring during deployment becomes more widely discussed. This will likely affect action-only subsection 3.1 monitorability, and potentially the propensity to verbalise relevant CoT reasoning subsection 2.1.2.

<sup>34</sup>One expert specifically identifies anti-jailbreak training as a likely driver of increasing evaluation awareness, citing evidence from the Opus 4.5 system card that features for identifying prompt injections increase over training.

<sup>35</sup>Three experts directly drew the details-vs-scenario distinction; two more gave reasoning consistent with it. No expert disagreed. One was somewhat more optimistic that scenario-driven awareness could also be tackled via better evaluations and beta testing.

<sup>36</sup>This issue will be particularly difficult for safety strategies depending on a lack of “deployment awareness” – where the model acts well in deployment because it retains uncertainty as to whether a deployment situation is actually a test or a training environment.

and so labs will be incentivised to reduce it, either by explicitly training against it or iterating model training and evaluation design against metrics of verbalised evaluation awareness. This may genuinely reduce evaluation gaming, but it also risks producing models that never explicitly mention evaluation awareness but still exhibit non-verbalised evaluation awareness and evaluation gaming (Schoen & Nitishinskaya, 2026).

### 3.2.3 Measuring evaluation awareness

**Attempt to measure the impact of evaluation gaming by intervening on evaluation awareness.** This involves performing interventions aimed at increasing or decreasing evaluation awareness, and observing how the model behaves differently in evaluations as a result. Approaches include resampling when verbalised evaluation awareness is present in the chain-of-thought (Schoen et al., 2025), or suppressing evaluation-awareness-related latent states via activation steering (Nguyen et al., 2025; Anthropic, 2025d; Hua et al., 2025). Divergences between the intervention and control conditions can provide evidence for the magnitude of evaluation gaming effects, but this evidence is generally weak as the interventions may not be targeted or effective. Conversely, if the evaluation gaming behaviour is known and specific, measuring that behaviour can serve as a measure of evaluation awareness, including non-verbalised awareness (Taylor, 2025; Kroiz et al., 2025).

**Measure verbalised evaluation (or oversight) awareness.** A simple approach (while chain-of-thought monitorability remains high) is to measure how frequently models verbalise beliefs about being evaluated in their chain-of-thought reasoning. This is the most common approach to identifying evaluation awareness in practice, e.g. Claude Sonnet 4.5 verbalised evaluation awareness over 80% of the time in alignment honeypot evaluations (Anthropic, 2025d). However, this approach becomes unreliable if models do not have a consistent propensity to verbalise evaluation awareness (see Section 2), as models are capable of modulating their behaviour in evaluations without explicit verbalisation (Taylor, 2025; Kroiz et al., 2025).

**Measure model capabilities for situational awareness.** Needham et al. (2025) and Laine et al. (2024) explicitly prompt models to classify whether transcripts come from evaluations or real-world deployment, finding substantial abilities at this task. Benchmarks such as these can be further developed and tracked over time to measure trends in situational awareness capabilities. Beyond distinguishing evaluation from deployment, it may be more important to understand whether models can deduce the purpose of the evaluation they are in, or accurately deduce details of how they are being trained or monitored.

**White-box techniques** can be developed to detect whether a model internally represents evaluation awareness. Nguyen et al. (2025); Anthropic (2025d) use probes and SAEs to detect evaluation awareness, and Anthropic (2025b, 2026); Fraser-Taliente et al. (2026) use natural-language autoencoders<sup>37</sup> to measure non-verbalised evaluation awareness (see subsection 5.1.2). White-box techniques can potentially provide a measurement channel that does not depend as heavily on chain-of-thought verbalisation. However these techniques need further development and validation.

### 3.2.4 Levers for preserving evaluation integrity

**Improve the realism of evaluations.** Mitigating evaluation awareness caused by unrealistic micro-details in the environment is likely to be tractable, whereas evaluation awareness caused by the nature of the scenario itself is highly likely to be more pernicious. Micro-details can be made more realistic by comparing to real deployment environments, e.g. minimally modifying

<sup>37</sup>specifically, the activation-verbaliser half of an NLA

production transcripts (OpenAI, 2025a) or agent scaffolding (Souly et al., 2025).<sup>38</sup> However, accurately recreating complicated production scaffolding from outside is often infeasible, which creates a structural asymmetry: the most realistic evaluations can only be run by labs with access to the full production stack. By contrast, scenario-driven evaluation awareness is when models deduce that they are in an evaluation because of the overall situation; perhaps they are placed in a high-stakes ethical dilemma that seems unlikely to occur in real life. Mitigating scenario-driven evaluation awareness may require more speculative approaches, such as activation steering to suppress evaluation-awareness representations.

**Production-based evaluations.** A particularly promising approach is to deploy a trusted model, collect deployment logs, filter for situations where sabotage opportunities arise, and then replay those situations with the untrusted model (OpenAI, 2025a). Because these evaluations are drawn from real deployment conversations, models exhibit substantially lower evaluation awareness than on traditional benchmarks. However, production-based evaluations cannot fully substitute for synthetic evaluations: they can be difficult to extend to long trajectories as the new model may explore differently to the old model, and they are limited to the distribution of situations that have already arisen in deployment (unless substantial edits are made, potentially increasing evaluation awareness).

**Beta testing as an alternative to evaluations.** Rather than constructing evaluation scenarios that a capable model might recognise as artificial, developers can run a beta test — a limited real deployment with consenting users and thorough monitoring — and use the resulting behaviour as evidence about how the model will act once more widely deployed. Because the model genuinely *is* deployed, the scenario-driven evaluation-awareness problem largely dissolves: there is no “tell” that the interaction is fake. However beta testing relies on sufficiently robust real-world monitoring, and also shares some limitations of production-based evaluations: access to detailed logs may be constrained in sensitive or proprietary deployment settings, and the beta test may fail to cover the full range of scenarios in which the model will eventually be deployed.

**Limit model affordances to reduce the surface for evaluation detection.** The more tools and information a model has access to, the more ways it can verify whether it is in an evaluation or deployment. For example, Claude Opus 4.6 used web search to identify the BrowseComp benchmark it was being evaluated on, and used Python access to decode the answer key (Anthropic, 2025a). Restricting affordances during evaluations (e.g. Python REPL access) reduces the surface for evaluation gaming. However, this is only effective if these affordances are also frequently restricted in deployment — otherwise, a model could simply condition its misaligned actions on having sufficient tool access to first verify it is not being evaluated.

**Activation steering to suppress evaluation awareness.** Activation steering can be used to suppress evaluation-awareness representations (Nguyen et al., 2025; Hua et al., 2025). Steering has shown promise at reducing verbalised evaluation awareness and eliciting higher levels of misalignment, but its effects are inconsistent and it does not seem to fully remove non-verbalised evaluation awareness (Anthropic, 2025b). This technique is currently immature and it is unclear how robust it will be as model introspection improves (McKenzie et al., 2026); nonetheless it may be among the best available approaches for mitigating pernicious forms of scenario-driven evaluation awareness.

---

<sup>38</sup>Three experts directly endorsed aligning eval details with real deployment; two more gave consistent reasoning. No expert disagreed. One cautioned that recreating complicated production scaffolding (e.g. Claude Code) from outside labs may become infeasible.

## 4 State, Memory, and Inter-Agent Communication (SMIA)

### Section Summary: State, Memory, and Inter-Agent Communication

|                             |  |
|-----------------------------|--|
| <b>Status:</b>              | Currently strong but evolving rapidly                        |
| <b>Risk of Degradation:</b> | ● High – medium-severity degradation likely                  |
| <b>Impact if Lost:</b>      | ● Medium – varying impacts, some severe                      |
| <b>Preservability:</b>      | ● Moderate – depends on architectural and deployment choices |

### Key Information

- **Continual learning poses a threat to auditing.** If model weights are updated during deployment, pre-deployment safety evaluations may no longer reflect how the model actually behaves, and the number of distinct model states requiring audit could proliferate rapidly.
- **State is stored in human-readable text.** Memories are stored as readable text (e.g. conversation history, markdown files), and multi-agent communication happens via text. These properties mean that memory and communications can be inspected and replayed for incident response, and that evaluators can construct realistic test scenarios by editing model context.
- **The most important levers are preserving text-based memory and communication if possible, and proactively developing interpretability tools for non-text alternatives.** Developers should also maintain comprehensive state logging where practical to support incident response and reproducibility.

This section discusses how models might change during deployment due to continual learning or modified state, and how instances of a model can communicate over time (which we think of as memory) or between parallel copies (inter-agent communication).

Early LLM assistants had no memory or stored state; every session would start from the same default prompt. Over the past year, model context windows have extended to tens of millions of tokens (Reid et al., 2024; Meta AI, 2025), and various memory systems have been introduced. Current assistants can refer to previous conversations, and coding agents store notes to their future selves in markdown files. For most frontier systems, model weights remain static from day to day, and no model developer has succeeded in unlocking true *general* continuous on-the-job learning. However, this is beginning to change: Cursor AI already uses online reinforcement learning to update its code completion model on a daily basis (Cursor AI, 2025a), and other labs regularly update model weights to patch jailbreaks. While current continuous-learning systems are narrow in scope, they demonstrate that the “weights are static” property is already eroding.

Multi-agent systems are under active development, with recent product releases of agent “team” or “swarm” features for coding. These teams are primarily a scaffold over existing agents, communicating between themselves via text much as humans communicate with agents. Future multi-agent systems may use more specialised communication channels, which could affect monitorability of these systems (Schroeder de Witt, 2025).

### 4.1 Safety-relevant properties

**Model weights are largely static during deployment.** For most frontier systems, pre-deployment model evaluations help us to predict how models will behave in deployment, both their propensities and their capabilities. If model weights are frequently updated during deployment, then pre-deployment evaluation results may not reflect the behaviour of the updated

model. Because weights are static, the number of distinct models requiring audit is small and does not proliferate with the number of tasks or deployments. Additionally, white-box monitoring techniques are trained assuming a fixed model, so if weights are regularly updated then these techniques may produce unreliable results unless they are also frequently updated to match.

**Pre-deployment evaluations can be designed to be representative of deployment behaviour.** While current deployment harnesses introduce some differentiation (e.g. custom instructions, information from previous chats), they can often be reproduced reasonably faithfully in testing. Model instances in deployment also have somewhat limited variation in capabilities and propensities between contexts, so pre-deployment evaluation results are often reasonably informative – though propensities do sometimes vary strongly with context (see [subsection A.3](#)).

**Memories and model state are stored in human-readable text, supporting honeypot construction and incident replay.** When monitoring detects misbehaviour, the full state can in principle be reconstructed and replayed for investigation. Evaluators can also edit or construct the model’s context merely by editing text, for example in order to run an alignment honeypot (although this is already limited by evaluation awareness; see [subsection 3.2](#)).<sup>39</sup>

**Communication and shared state between models in multi-agent settings is human-readable.** Instances communicate (both directly and indirectly) via human-readable text, so we can monitor whether a model is trying to jailbreak another, instructing it to take harmful actions, or coordinating on subverting oversight.<sup>40</sup> Communication can happen between parallel running instances, or sequentially from one instance to the next, e.g. via compaction messages or cues left in the environment.

## 4.2 Degradation pathways

| Oversight Degradation Pathways                              |                                   |                  |                       |          |
|---|-----------------------------------|------------------|-----------------------|----------|
| Mechanism   | Properties affected               | Timeline         | Likelihood            | Severity |
| <a href="#">Weight-based continual learning</a>             | Weights are static                | Medium-term      | Likely                | ●        |
| <a href="#">Longer contexts and advanced memory systems</a> | Representativeness of evaluations | Gradual, ongoing | Highly likely         | ●        |
| <a href="#">Vector-based memory</a>                         | Context is readable               | Medium-term      | Realistic possibility | ●        |
| <a href="#">Vector-based inter-agent communication</a>      | Communication is readable         | Medium-term      | Realistic possibility | ●        |

**Model weights are likely to be regularly updated during deployment (continual learning), meaning pre-deployment evaluations may stop being representative of deployment behaviour.** Continual learning could provide significant gains in real-world utility of AI systems, and some developers are already moving in this direction.<sup>41</sup> This may take the

<sup>39</sup>AI-specific logging should also be supported by more traditional cybersecurity logging and tracking of the surrounding code and events which are executed.

<sup>40</sup>Three experts identified text-based inter-agent communication as a current safety property of multi-agent systems: one called it “*a nice emergent property that human-readable text is used*”; one called bureaucratic text-based agent systems “*easier for us to monitor than if it’s one horrific end-to-end system*”; and one described file-system and task-tracker communication as “*almost like a version of chain of thought where models have to verbalise something*”.

<sup>41</sup>For example, Cursor AI uses online reinforcement learning to update its code completion model throughout the day ([Cursor AI, 2025b](#)).

form of frequent learning from online data (Shi et al., 2024), an updating memory module (Behrouz et al., 2025), updating the model weights via context distillation (Snell et al., 2022; Kujanpää et al., 2024), or even *test-time training*, where weights are updated during inference itself (Tandon et al., 2025). Updates to model weights shared between all users could mean evaluation results become out of date as models change over time, while per-user updates would lead to a hard-to-track diversity of behaviour among models in deployment. The frequency of weight updates matters: if developers update weights monthly, periodic re-auditing may remain feasible; if updates occur daily or continuously, the cost of keeping evaluations current could become prohibitive. Unlike traditional software, where a targeted code change typically affects a limited and well-understood set of behaviours, weight updates can have diffuse and hard-to-predict effects across the model’s behaviour, substantially raising the cost of re-auditing relative to standard software-engineering assumptions.<sup>42</sup>

**Context lengths are highly likely to get longer, enabling more in-context learning and making model behaviour more dependent on an increasingly complex state.** Larger and more complex state is harder to fake, making it harder to audit models by putting them in artificially created scenarios. Similarly, if monitoring detects an instance of misbehaviour, it could be harder to replicate the same state in order to study the reasons for the model’s actions.<sup>43</sup>

**Memories and state are increasingly shared between multiple instances of a model** – this can happen either intentionally, e.g. through shared files documenting acquired skills, or unintentionally, e.g. through text online left by previous models.<sup>44</sup> If unmonitored this could lead to collusion between models. Spread and amplification of misaligned memories could be hard to predict in advance, and unlikely to surface in limited pre-deployment testing (Mallen, 2025).<sup>45</sup>

**Developers may switch to vector-based approaches to memory, reducing the ability to monitor and edit a model’s memory.** Note that we distinguish this from retrieval-augmented generation (RAG), which stores and retrieves text documents and is already widely used. The concern here is with systems that store memories as opaque vector representations (e.g. KV cache pairs) rather than human-readable text. This could happen in parallel with models switching to latent reasoning (see Section 2), or separately – see Yang et al. (2024) for an example of such an approach. Vector-based memory would be much harder to monitor or edit than text-based memory, but perhaps easier than weight-based continual learning. The monitorability of memory systems varies along three dimensions: how interpretable is the format in which memories are stored, how interpretable is the retrieval mechanism, and how interpretable is the effect of memory edits on model behaviour?<sup>46</sup>

<sup>42</sup>One expert noted that the capability case for weight-based continual learning is the distillation of explicit (“*system 2*”) reasoning into instinctive (“*system 1*”) responses, making the most useful continual-learning regimes also the least interpretable.

<sup>43</sup>Three experts noted that very long contexts and trajectories already strain monitoring: one observed that chains of thought of tens of thousands of tokens are very hard for humans to parse; another warned of days-long agentic deployments; and a third noted that “*almost all interpretability methods today don’t scale up*” to long generations.

<sup>44</sup>This is already becoming a challenge with current models; for example Anthropic found that when running the BrowseComp web search benchmark, Claude Opus 4.6 would observe traces left behind by previous runs (Coleman, 2026).

<sup>45</sup>At least four experts identified shared memory as a channel through which agents could coordinate or collude: writing to shared memory was described as functionally equivalent to message-passing between agents, with one expert noting that with continuous-representation memory it would be “*really rough to uncover what’s going on there*”, and another warning that a strategy discovered by one agent could “*quickly percolate around*” connected instances. Two further experts gave indirect support, describing how environmental state (files, logs) lets agents leave “*a trail of breadcrumbs*” for future instances.

<sup>46</sup>Four experts flagged vector or KV-cache style memory as the most likely near-term opaque-memory pathway: one called natural-language memory “*way less concerning than activations we don’t understand*”; another foresaw

**Multi-agent systems may use vector-based communication, reducing the ability to monitor inter-model communication.** As with memory, a switch to latent reasoning may naturally lead to models communicating directly via latent representations, or non-text communication could be adopted separately (e.g. building on approaches such as [Ye et al. \(2025\)](#) that share KV cache data between agents). Similarly to vector-based memories, this form of communication would be much harder to monitor. If agents communicate via shared vector memory stores, misaligned propensities or strategies discovered by one agent could in principle rapidly propagate to connected agents with little oversight, creating what amounts to an unmonitorable “hive-mind” ([Schroeder de Witt, 2025](#)).<sup>4748</sup>

**Privacy and access control in multi-agent settings need careful design to avoid creating unmonitorable blind spots.** In settings where models are owned by different actors, privacy constraints can mean that no single human has permission to inspect all inter-agent communications.

### 4.3 Measuring SMIA oversight

- **Model developers should disclose when they use techniques which could degrade oversight.** Weight-based continual learning, vector memory and vector communication are discrete architectural details: the development and introduction of these approaches is feasible to track if developers are transparent about their use.
- **The extent to which capabilities and propensities drift depending on context should be investigated.** Propensity evaluations should stress-test alignment across long contexts intended to shift their persona, and techniques such as persona tracking ([Lu et al., 2026](#)) or logit-diff amplification ([Aranguri & McGrath, 2025](#)) should be investigated to find signs of these shifts.
- **Multi-agent monitorability benchmarks should be constructed.** The ability for models to collude or pass unmonitored signal should be measured, and similarly for the monitorability of memories, extending existing research in the area of AI Control.
- **The monitorability of different memory architectures should be evaluated comparatively.** As new memory and learning systems are introduced, evaluations should measure how effectively each architecture supports inspection, editing, and replay of stored state — key requirements for auditing and incident response.

### 4.4 Levers for preserving SMIA oversight

**Developers can best preserve monitoring of memory and communication by continuing to use plain text, unless they can maintain equivalent monitorability for non-text systems.** Developers should be transparent about what systems are being used, and allow external auditors to review what is stored in memory during audits. Developers should only adopt non-text systems if they can make safety cases for how they will mitigate the risks. For example, if weight-based continual learning is adopted, developers should explain how they will

---

shared memory banks replacing per-session memory wipes for efficiency; a third warned that vector-based memory “*would make it easier to subvert things*”; and a fourth articulated the three-axis taxonomy adopted in this paragraph (interpretability of storage, retrieval, and editing).

<sup>47</sup>Two experts described this specific pathway: one warned that “*if they figure out a way to do cache-to-cache communication efficiently... you just completely lose that avenue of monitorability*”; another argued that combining neuralese reasoning with a shared neuralese memory store would push AI systems toward “*a hive mind where they can all communicate to each other constantly*”.

<sup>48</sup>Models could also collude via steganography – subtle patterns in ostensibly natural text ([Mathew et al., 2024](#)). See [section 2.2.3](#) for more discussion on steganography.

keep pre-deployment evaluations relevant, and how they will ensure that white-box monitoring techniques remain effective.

**For multi-agent systems, maintaining text-based inter-agent communication is the single most important lever.** The current default of text-based communication between agents provides a natural monitoring surface. Where agents share a persistent memory store, the contents of that store should be human-readable and auditable. Developers should also consider using diverse model families for monitoring setups, to reduce the risk of collusion arising from shared representational structure (Bostock et al., 2026).

**Interpretability tools for non-text memory and communication should be proactively developed.** It may be possible to repurpose existing techniques for interpreting model activations.

**To preserve our ability to reconstruct and find root causes of misbehaviour detected by monitoring, we should ensure that we log the full state needed to reconstruct an interaction,** including the state that models store outside themselves (e.g. writing markdown files in the filesystem). Although some aspects of a model’s environment may be impractical to log in full, and capturing consistent state across distributed systems is a challenge even in regular software, the ability to reconstruct and replay incidents is a key part of being able to draw lessons from them, and to build realistic test scenarios. Tools for editing and faking state should also be further developed, to support evaluations and honeypots.<sup>49</sup>

**When considering adopting weight-based continual learning, developers should anticipate how it might make safety cases based on pre-deployment evaluations more difficult.** In the event that weight-based continual learning is adopted, we should develop interpretability techniques that can track the evolution of models over time as they update. If interpretability tools such as probes do not also update as model parameters change, previously valid monitoring results may become unreliable.

---

<sup>49</sup>Two experts argued that full state logging and replay are critical for incident response: one called for logging “*not just what action the AI did, but the whole context and history that led up to it*” for full reproducibility; another warned that reproducing incidents is “*becoming way harder over time*” as models operate in complex scaffolded environments.

## 5 White-Box Oversight

### Section Summary: White-Box Oversight

|                             |  |
|-----------------------------|--|
| <b>Status:</b>              | Varies widely: probes production-ready; SAEs used for auditing; ambitious interp research-stage. |
| <b>Risk of Degradation:</b> | ● Medium – Representational drift and dependence on CoT legibility are real risks.               |
| <b>Impact if Lost:</b>      | ● Moderate – Provides independent signals and a backstop if other methods degrade.               |
| <b>Preservability:</b>      | ● High – Plenty of scope to improve white-box methods.   |

### Key Information

- **White-box monitorability and auditing techniques vary widely on maturity and usefulness, but are largely improving.**
- **White-box techniques currently benefit from human-interpretable representations.** There is a realistic possibility that representations will become less interpretable due to increased post-training, making white-box oversight more difficult.
- **White-box techniques currently benefit from chain-of-thought legibility.** If CoT becomes less human-legible, white-box interpretations will become more difficult to ground.

**White-box methods operate on models’ internal states, and so are less vulnerable to the model choosing which information to reveal in its outputs or reasoning** (Casper et al., 2024). The model’s underlying computation routes through its activations, and white-box methods aim to read this information directly. This makes white-box monitorability a complement to other monitoring approaches, and a backstop if they degrade. However this depends on model internals being structured in ways that available techniques can read and interpret — for example, that safety-relevant concepts are represented in linearly accessible directions, and that internal representations remain sufficiently human-aligned for probes and interpretability tools to detect them.

**White-box methods span a wide maturity spectrum.** Probes (lightweight supervised classifiers trained on model activations) have been integrated into some production settings for specific use cases and are cheap enough for real-time monitoring (Kramár et al., 2026) (see Section 5.1.1). Unsupervised interpretability approaches such as sparse autoencoders (SAEs) are used experimentally for pre-deployment auditing at Anthropic (Anthropic, 2025b) (see Section 5.1.3). By contrast ambitious mechanistic interpretability (aiming towards full model understanding, perhaps achieved by circuit analysis or parameter decomposition) remains primarily a research endeavour (see Section 5.1.4). The key question is whether these methods can mature fast enough to compensate for degradation in other monitoring properties. We discuss challenges to white-box oversight in Section 5.2, measuring white-box oversight in Section 5.3, and levers for preservation and improvement in Section 5.4.

### 5.1 Relevance to safety claims

#### 5.1.1 Probes

**Probes are supervised classifiers trained on model activations to detect specific properties** (Alain & Bengio, 2016; Belinkov, 2022). Probes have shown promise at detecting deception (Goldowsky-Dill et al., 2025), cyber-offensive misuse (Kramár et al., 2026), dangerous

goals (MacDiarmid et al., 2024), high-stakes situations (McKenzie et al., 2025), and hallucinations (Obeso et al., 2025). These results demonstrate that safety-relevant information is often accessible in model activations, although generalisation to new settings is far from automatic (see below) and probes are not always robust enough to be relied upon alone (Taylor et al., 2025; Goldowsky-Dill et al., 2025). Related techniques include steering vectors for controlling model behaviour (Subramani et al., 2022; Turner et al., 2024; Zou et al., 2023), and latent-space anomaly detection (Podolskiy et al., 2022; Johnston et al., 2025).<sup>50</sup>

**Cost is the decisive advantage of probes for production monitoring.** Probes are lightweight classifiers, adding negligible overhead in terms of latency or computational cost. Low cost is important for broad-coverage monitoring, and low latency is important for domains where monitoring must be done online rather than delayed, e.g. to prevent data exfiltration. This kind of low-cost, broad-coverage, low-latency monitoring generally either requires deploying a probe, or deploying a small language model, and probes leveraging the representations of large models can often be more robust than small language models which only have access to text.<sup>51</sup>

**Probes fit naturally into hierarchical monitoring schemes.** Probes are cheap and fast, so they can naturally sit at the beginning of an escalation hierarchy, where suspicious cases are escalated to more expensive reasoning models or richer interpretability tools. For example, Constitutional Classifiers++ (Cunningham et al., 2025) demonstrate this kind of multi-tier approach. Probes can still be valuable even if they have a high false positive rate, assuming more expensive but more reliable techniques exist to validate these cases.<sup>52</sup>

**Probe generalisation is sometimes patchy, and probes are hard to red-team.** Probes do often generalise well for a few concepts that appear naturally and frequently in pre-training, or which are directly evident in the text itself (MacDiarmid et al., 2024; Kirch et al., 2025; Goldowsky-Dill et al., 2025), but some experts noted difficulties replicating deception-probe results on other models, and probes sometimes generalise too poorly across distribution shifts to be practical without substantial labelled data. When validating deception detectors, the primary bottleneck is labelling and good evaluations: e.g. it is unclear what counts as “deceptive” model behaviour in many settings (Smith et al., 2025). Red-teaming and stress-testing probes can also be more conceptually difficult than red-teaming standard text-based monitors.

### 5.1.2 Activation oracles and natural-language autoencoders

**This subsection covers methods that translate internal activations into natural-language descriptions.** Approaches in this family differ in whether they are trained in a supervised or unsupervised way, and in whether a single model is trained to handle arbitrary questions about an activation or to produce a free-text explanation of it. We treat three lineages in turn: *activation oracles* (supervised QA), *self-explanation / introspection training* (training a model to explain its own computations), and *natural-language autoencoders* (unsupervised free-text explanations).

**Activation oracles are supervised models trained to answer natural-language questions about activations** (Karvonen et al., 2025), building on earlier work such as LatentQA (Pan et al., 2026). Given a labelled corpus of (activation, question, answer) tuples, an activation oracle can generalise to answer out-of-distribution questions about activations. The approach is agnostic to which model is used as the explainer.

<sup>50</sup>Cyber-offensive misuse is likely to be a particularly difficult domain for probes, as many cyber techniques are dual-use, and cyber-attackers are typically used to navigating past blockers creatively.

<sup>51</sup>Two experts emphasised the cost-efficiency of probes as a large advantage for production monitoring.

<sup>52</sup>Two experts independently described this hierarchical/escalation pattern as their preferred operating model: one called it “*my intuition for what an ideal control scheme would look like*”.

**Self-explanation / introspection training** (Li et al., 2025) trains the target model itself to explain its own computations, including aspects beyond just activations, aiming to make reasoning traces and self-reports more faithfully reflect internal computations. Early evidence points to a *privileged self-access* effect: models explain their own computations more accurately than external explainer models can.<sup>53</sup>

**Natural-language autoencoders (NLAs) produce free-text explanations of activations in an unsupervised way** (Fraser-Taliente et al., 2026). An NLA consists of two jointly trained components: an *activation verbaliser* (AV), which maps an activation to a natural-language description, and an *activation reconstructor* (AR), which maps the description back to a reconstructed activation. The pair is trained via reinforcement learning to minimise reconstruction error. At Anthropic, NLAs have been used in the pre-deployment audits of Claude Opus 4.6 and Claude Mythos Preview, including to surface non-verbalised evaluation awareness and grader-awareness in training transcripts.<sup>54</sup>

**These approaches are promising but immature.** Many experts express excitement about this family of techniques, while stressing that it is a nascent research area.<sup>55</sup> Jakkli et al. (2026) find that current activation oracles are difficult to use reliably on safety-relevant tasks, with responses that can be vague, hallucinated, or driven by surface text patterns. The NLA paper also lists many limitations, including confabulation, lack of mechanistic grounding, excessive expressivity of the activation verbaliser, and high training and inference cost (Fraser-Taliente et al., 2026). Currently these methods are more effective at raising hypotheses than robustly verifying them.

**The explainer’s own cognition could bias its explanations.** Because the explainer is itself an intelligent model, it performs additional processing which can contaminate its descriptions of the subject model’s activations. For example, deceptive models could deliberately lie when explaining their own activations in certain circumstances outside the training distribution.<sup>56</sup>

### 5.1.3 Sparse Autoencoders (SAEs)

**Sparse autoencoders (SAEs) aim to decompose model activations into interpretable features**, providing an unsupervised window into what concepts a model is representing (Sharkey et al., 2022; Cunningham et al., 2023; Bricken et al., 2023). Related unsupervised approaches include natural-language autoencoders (see subsection 5.1.2); codebook features (Tamkin et al., 2024); mixture-of-factor-analyzers (Shafran et al., 2026); semi-nonnegative matrix factorisation (Shafran et al., 2025), transcoders, which map between layers (Dunefsky et al., 2024; Ameisen et al., 2025); and model diffing techniques that identify behavioural changes introduced by fine-tuning (Lindsey et al., 2024; Aranguri & McGrath, 2025).

**SAEs are useful for open-ended debugging, but struggle to compete with simpler methods on narrowly targeted tasks.** SAEs do not require the analyst to specify what to look for in advance, making them most valuable for open-ended debugging and surfacing

<sup>53</sup>One expert also theorises that training models to explain their own internal computations in this way would also shape the internal computations themselves to be simpler and easier to explain, rather than merely shaping the model’s explanations of them.

<sup>54</sup>This is the technique originally referred to as “an unpublished variant of activation oracles” in the Opus 4.6 system card (Anthropic, 2025b, 2026; Fraser-Taliente et al., 2026).

<sup>55</sup>Three experts singled out activation oracles, natural-language autoencoders, or related decoding approaches as a particularly exciting direction — one calling it the “most exciting” route to addressing neuralese — while emphasising that the area is nascent and not yet reliable enough to depend on.

<sup>56</sup>This is a particular concern for measuring evaluation awareness: the explainer itself may be aware of the evaluation context, making it difficult to distinguish whether awareness is encoded in the subject model’s activations or is a product of the explainer’s own cognition. Despite this limitation, NLAs have been used for evaluation awareness assessment in practice (Anthropic, 2025b).

unknown unknowns (Marks et al., 2025; Anthropic, 2025b). However Nanda et al. (2025) notes that open-source SAEs often underperform simpler baselines on downstream tasks (Chaudhary & Geiger, 2024; Huang et al., 2024). Additionally, SAE features learned from identical data can vary substantially across random seeds (Paulo & Belrose, 2025), and SAE architectures encode implicit assumptions (Hindupur et al., 2025; Menon et al., 2025). None of these results show SAEs are fundamentally broken, and most of the cited papers propose fixes rather than abandonment; but they are reasons not to over-rely on SAE features as ground truth.

#### 5.1.4 Ambitious interpretability

**Ambitious mechanistic interpretability (AMI) aims to go beyond detecting specific features and understand *how* models compute their outputs** (Gao, 2025). Building on earlier circuit analysis methods (Olah et al., 2020; Elhage et al., 2021; Wang et al., 2023), attribution graphs trace how information flows from inputs to outputs (Ameisen et al., 2025). Applied to Claude 3.5 Haiku, they have revealed e.g. forward planning in poetry, and some mechanisms underlying hallucination and jailbreak vulnerabilities (Lindsey et al., 2025). Weight-space methods decompose parameters directly rather than relying on activations (Braun et al., 2025), while complementary approaches study how computational structure emerges during training (Hoogland et al., 2025). Although not yet mature, these more ambitious approaches could in principle support much stronger safety claims than probes or SAEs – e.g. that the model’s computation is as-desired, rather than merely that no malign features were detected.

##### Expert Disagreement

**Experts disagree on the prospects of ambitious interpretability.** Some experts see the prospects of ambitious interpretability (as well as SAEs and activation oracles to a lesser extent) as an open research question, highlighting that it is not clear whether they would ever be centrally useful in practice. Others draw optimism from recent successes of interpretability, e.g. in unlearning (Gur-Arieh et al., 2025; Cloud et al., 2024; Sheshadri et al., 2025), or from the possibility of interpretability being useful for model capabilities, so that oversight follows as a side effect of capability investment into e.g. decomposable architectures.

## 5.2 Challenges for white-box oversight

Perhaps the most fundamental risk is simply that many **white-box oversight methods are still not fully mature, and could fail to improve fast enough to serve as dependable, legible evidence** before other monitoring avenues degrade. However there are also many factors which could increase the difficulty of white-box oversight:

##### Oversight Degradation Pathways - White-Box Methods

| Degradation pathway                               | Timeline                | Likelihood            | Severity |
|---|-------------------------|-----------------------|----------|
| <b>Representational drift</b>                     | Medium-term             | Realistic possibility | ●        |
| <b>Black-box degradation undermines grounding</b> | Follows CoT degradation | Likely                | ●        |
| <b>Scaling to long reasoning traces</b>           | Current limitation      | Ongoing               | ●        |
| <b>Non-linear representations</b>                 | Long-term               | Unlikely in near term | ●        |
| <b>Learned resistance to interpretability</b>     | Medium-term             | Realistic possibility | ●        |

|   |             |                       |   |
|---|-------------|-----------------------|---|
| <b>Situational awareness of white-box interventions</b> | Medium-term | Realistic possibility | ● |
| <b>Practical barriers to adoption</b>                   | Ongoing     | Current limitation    | ● |

### 5.2.1 Representational drift

Many current white-box methods rely on the fact that safety-relevant concepts remain connected to human-interpretable concepts from pre-training. Probing and interpretability can be significantly complicated if the labels a probe was trained on cease to line up with the intended part of the model’s internal state. This can happen in several distinct ways, which have different implications for mitigation.

The most empirically grounded concern is *belief drift* rather than representational drift proper. Cundy & Gleave (2025) find that models which have been heavily RL-optimised to perform a certain behaviour can continue to have human-interpretable representations of, e.g., deception, but shift their beliefs about what falls under those concepts – e.g. from “hardcoding tests is generally bad” to “hardcoding is usually fine.” The probe is still picking up on the model’s concept of deception; the model has just stopped classifying the target behaviour as deceptive.

A separate concern is subtle drift in what the representation tracks. A model’s internal representation might shift from tracking “this is actually true” to tracking “my overseers believe this is true” more strongly, causing a truth probe training to select the latter representation rather than the former, leading to catastrophic failures where they differ.

A further, more speculative concern is drift into entirely alien concepts which fail to map cleanly onto the human ontology that probes and interpretability tools were built around.

#### Expert Disagreement: Alien Representations

**Experts disagree on the impact of models developing fundamentally non-human representations.** As models become more capable than humans, some experts expect ontologies to shift away from human-interpretable concepts, undermining interpretability and oversight. Four experts raised this concern, citing the gap between current interpretability tooling (mostly anchored to human concepts) and how superhuman models may factorise knowledge. Other experts are uncertain, noting that larger and more capable models could have more unified representations (aiding interpretability) or more specialised non-human concepts (undermining it).<sup>a</sup> Finally, others think the impact of truly non-human concepts is likely to be limited — expecting the most safety-relevant concepts (e.g. deception, misaligned goals) to be human-understandable assuming they can be accurately found.<sup>b</sup>

<sup>a</sup>One expert explicitly articulated both sides of this trade-off.

<sup>b</sup>Two experts raised this prediction.

**Empirically, pre-training representations may not persist as reinforcement learning (RL) scales.** There is a realistic possibility that important concepts will become untethered from the pre-training distribution as more learning happens during RL, undermining white-box oversight. Concerningly, there is evidence that representational drift during RL can make even freshly trained deception probes less effective (Taufeeque et al., 2026). Although there is not yet significant evidence of this occurring in production despite a large scale-up in RL over 2025, it is a relatively underexplored issue.<sup>57</sup>

<sup>57</sup>Two experts flagged this as a key open question: one observed empirically that probes trained on a pre-RL

### 5.2.2 Black-box degradation harms white-box oversight

**White-box methods are almost always *grounded in the inputs and outputs of the model***, even though the methods themselves operate on internal states. For example, SAE features are typically interpreted by analysing the distribution of tokens on which the feature fires. Likewise, probes are trained by constructing token distributions on which the probe should or should not fire.

**If CoT becomes less human-legible, interpretations will become more difficult to ground.** If model inputs or outputs become more opaque, making sense of internal features also becomes harder – latents must be traced further to understand the behaviours they represent. Probes also benefit from picking up on verbalised CoT signals: when model organisms were created that exhibit deceptive behaviour without producing textual evidence, probes performed substantially worse (Boxo et al., 2025).

### 5.2.3 Scaling challenges

**Scaling to long contexts and reasoning traces is the field’s key near-term challenge, though likely not a fundamental one.** Most interpretability methods still focus on a single forward pass, while reasoning models distribute computation across many tokens. Long structured reasoning and complex multi-agent scaffolding create challenges with knowing where to look, or how to localise computations.<sup>58</sup> This limitation is best seen as an under-invested area rather than a fundamental barrier. Promising research directions exist, such as tracking representations over context, attention probes (Kramár et al., 2026), activation oracles, but are nascent.<sup>59</sup>

### 5.2.4 Non-linear representations

**Some methods assume that relevant concepts use linear structure in activation space** — notably linear probes, SAE feature directions, and steering-vectors. Many white-box methods do not rely on this assumption: for example, non-linear probes, activation oracles, natural-language autoencoders (subsection 5.1.2), meta-models of activation space (Luo et al., 2026), and parameter decomposition techniques (Braun et al., 2025) do not assume linearity. So failure of the *Linear Representation Hypothesis* (LRH) would shift the portfolio of interpretability tools rather than eliminate white-box oversight. The central concern is therefore not that the LRH might fail, but that it might fail *silently* — with linear methods appearing to work while systematically missing safety-relevant structure. Detecting such a failure requires regularly comparing linear methods against non-linear alternatives. The status of the LRH itself is contested among interviewed experts; see the expert-disagreement box below.

#### Expert Disagreement: the Linear Representation Hypothesis

**Experts disagree on the status of the Linear Representation Hypothesis** — the claim that models encode concepts along linear directions in activation space. Some experts treat approximate linearity as a useful working assumption for developing practical interpretability tooling. Although some features are known to be *irreducibly multi-dimensional*

model degraded after RL even without adversarial pressure (attributing this to representational drift); another noted that “*the more of your learning happens during RL, the less tethered it is to the pre-training distribution and the more it could be exotic or alien*”, adding that “people haven’t really thought about this problem much”

<sup>58</sup>Some more computationally intensive methods (especially ambitious approaches) can also be expensive to scale up to large models.

<sup>59</sup>Two experts identified scaling interpretability to long reasoning traces and complex scaffolds as a central near-term challenge: one called it the biggest gap (“*almost all interpretability methods today don’t scale up*”); another raised the basic “*what do you even look at?*” problem in multi-agent scaffolds.

– for example, days of the week are encoded as a 2D circle (Engels et al., 2024) – these still live in linear subspaces and are often accessible to linear methods; whether models also encode safety-relevant structure in truly non-linear ways remains an open question (Gorton, 2026). Others argue the hypothesis is either empirically wrong or conceptually ill-formed, pointing to non-linear probing that outperforms linear probing (Kirch et al., 2024), the use of low-rank (rather than strictly linear) fine-tuning adapters, polytope-based accounts of model internals (Black et al., 2022), steering and monitoring methods that work without linear feature directions (Beaglehole et al., 2025), and conceptual work arguing that the LRH is not well-defined. Some experts argue that nonlinear or parameter-based alternatives (subsection 5.1.3) should replace linear probes and SAEs.

### 5.2.5 Learned resistance to interpretability

**White-box methods seem relatively robust to RL pressure, but training against white-box methods is still risky.** As with other forms of oversight, using white-box methods to detect and train against undesired behaviour opens up the risk that the model learns to subvert the monitor (Sharkey, 2022). However there are early signals that newly-trained probes are not as vulnerable to this as CoT monitors: Taufeeque et al. (2026) observe that penalising probe detections during RL does not increase new probe subversion beyond what occurs from baseline representational drift, Cundy & Gleave (2025) observe that white-box monitoring can be relatively robust to some forms of training pressure (but not others), and Gupta & Jenner (2025) find that monitors targeting specific activation positions are vulnerable to RL-based evasion, whereas monitors which aggregate over many tokens remain substantially more robust. Overall, if white-box methods are used during training, the training process should be stress-tested and validated with care.

### 5.2.6 Situational awareness

**Models may become increasingly aware of white-box interventions and act to reduce their effects** (Sharkey, 2022). There is emerging evidence that models can resist white-box manipulation. Recent models can sometimes detect (Lindsey, 2025) and spontaneously resist (McKenzie et al., 2026) activation steering in-context. Recent work on introspection (Lindsey, 2025) also suggests that frontier models can to some extent *control* their own internal states. However spontaneously subverting white-box monitoring would likely require more situational awareness of when and how the monitoring is being applied, and more systematic in-context control over internal states than has been observed to date. While models can be trained to hide their thoughts from activation monitors when told about the monitor in the prompt (McGuinness et al., 2025; Bailey et al., 2024), this is unlikely to arise naturally in the near term without explicit optimisation pressure.

### 5.2.7 Practical issues

**There is a risk that white-box techniques are not implemented widely due to practical difficulties.** White-box techniques require access to more state than is typically provided over APIs, so new APIs are necessary for external actors to access white-box monitors. This also increases friction, as the model provider may need to be responsible for training and deploying the monitor before serving the result. Production inference codebases are highly optimised, and at least one frontier lab found it took significant effort to deploy even simple probes due to this (Kramár et al., 2026).

### 5.3 Measuring white-box oversight

**Study real models in depth.** One route to validation is to study models carefully in realistic situations, to build mechanistic accounts and judgements of whether white-box approaches are tracking the relevant structure.

**Use model organisms for stress-testing with known ground truth.** Better and more realistic model organisms with ground-truth signals are key for studying whether white-box oversight is robust (Hubinger, 2023). Auditing games – planting a hidden objective and testing whether methods can surface it (Marks et al., 2025; Taylor et al., 2025) – provide structured test frameworks where the ground truth is known. The open challenge is judging whether model organisms are realistic models of future threats.

**Improve benchmarks for white-box monitors.** Unlike other methods, access to model internals is needed to benchmark white-box monitors. Kretschmar et al. (2026) provides one benchmark for detecting lies in open-weight models; more such benchmarks, especially ones with hard-to-fake downstream tasks, would be valuable. Comparing the results to other approaches (e.g. CoT or action monitoring) can provide strong evidence that the monitor is extracting real additional signal.

**Track representational drift and out-of-distribution transfer.** Developers should measure how well probes and SAEs trained on the base model transfer to the post-RL model, and more generally whether probes trained on human behaviour or off-policy text continue to transfer to on-policy behaviours (Kirch et al., 2025), or transfer across domains (Goldowsky-Dill et al., 2025).

### 5.4 Levers for improving white-box oversight

**Continue to improve white-box techniques,** across the full portfolio from probes to ambitious interpretability. Probes are production-ready but can struggle with complex aggregation over long contexts, so moving to approaches such as attention or transformer-based probes may be an improvement (Kramár et al., 2026). Obtaining reliable lie-detection techniques would be particularly high-leverage against risks from coherently misaligned models. Activation oracles, natural-language autoencoders, and introspection training are promising new directions (Section 5.1.2) that require further work and testing. More ambitious approaches require the most work, but may be correspondingly valuable.

**Encourage early integration into production.** The longer probes are absent from the inference stack, the more costly integration becomes (Kramár et al., 2026). Starting with currently pressing and tractable use cases (e.g. misuse detection) gets the infrastructure in place, after which adding further safety probes is cheap. Online probe retraining – where probe firings are escalated for labelling, then used to retrain – could mitigate representational drift and distribution shift without requiring probes to generalise perfectly from the outset.<sup>60</sup>

**Using interpretability for capabilities could create monitorability by default.** If labs find interpretability practically useful for training – e.g. for removing sycophancy via persona vectors, precise unlearning (Gur-Arieh et al., 2025; Cloud et al., 2024; Sheshadri et al., 2025), or avoiding emergent misalignment – they may integrate it into standard workflows, making oversight infrastructure a side effect of routine engineering rather than an additional safety tax. However, experts disagree on whether this path is desirable: some see it as the most promising route to widespread adoption, while others caution that interpretability advances which also improve capabilities may on net be harmful by accelerating the timeline for dangerous systems, leaving less time for safety work to catch up.

---

<sup>60</sup>One expert explicitly raised this online retraining mechanism.

## 6 Training for Oversight

### Section Summary: Training for Oversight

|                             |   |
|-----------------------------|---|
| <b>Status:</b>              | Many approaches under active development  |
| <b>Risk of Degradation:</b> | ● Moderate – adversarial generalisation failure almost certain to become harder to rule out as models grow more capable |
| <b>Impact if Lost:</b>      | ● High – training-based approaches are a key fallback if other monitoring surfaces (e.g. CoT necessity) degrade         |
| <b>Preservability:</b>      | ● Moderate – significant room for improvement but fundamental challenges around generalisation and training data remain |

### Key Information

- **Training the *untrusted* model itself — for honesty, legibility, confession, or faithfulness — is a potentially promising fallback if CoT monitorability degrades**, but is most exposed to *adversarial generalisation failure* (a capable misaligned model behaving well in training while continuing to misbehave in deployment).
- **Training a *separate* monitor model — e.g. a CoT monitor, supervised probe, or unsupervised interpretability tool — sidesteps most adversarial-generalisation risks**, but faces exploitability, weakness, and non-adversarial generalisation challenges.
- **A key cross-cutting issue is where the training data comes from.** Approaches must either source training data that covers the actual misalignment they need to catch, or rely on generalisation from synthetic or proxy data. Both ends of this spectrum have challenges.
- **Expert attitudes towards training for oversight span a wide range.** Some experts expect untrusted model training to generalise in the intended ways by default (or at least to constrain plausible misalignment); others argue that untrusted model training approaches are too vulnerable to oversight awareness and other systematic failures that arise alongside sufficient model capability.

Many oversight approaches hinge on training – either training the model itself to be honest or understandable, or training a separate model to oversee it. This section focuses on training as a tool for *building or preserving* monitorability, rather than the processes by which training may *degrade* monitorability (e.g. training against monitors in a bid to remove caught behaviours, as discussed in [section 2](#) and throughout the report).

## 6.1 Enhancing oversight through training interventions

Experts identified training-based approaches as a key area for enhancing or preserving monitorability, particularly as a natural next step if chain-of-thought monitorability degrades. There are a variety of approaches, and most are new or still developing:

| Training Interventions for Oversight        |   |   |
|---|---|---|
| Training Approach                           | Model Trained                           | Description   |
| <b>Honest character training</b>            | Untrusted model itself                  | Training models for a broad disposition towards honesty (Askell et al., 2026; Maiya et al., 2025).  |
| <b>Confession / self-report training</b>    | Untrusted model itself                  | Training models to self-report their own misaligned behaviours or hidden objectives, either when prompted (Joglekar et al., 2025; Li et al., 2026) or without being prompted (Lee et al., 2026; Taylor, 2026; Kirk et al., 2026; Anthropic, 2025c). |
| <b>Verbalisation propensity training</b>    | Untrusted model itself                  | Training models to verbalise safety-relevant parts of their reasoning (e.g. reward hacking) unprompted (Turpin et al., 2025).   |
| <b>Legibility training</b>                  | Untrusted model itself                  | Training models so their reasoning traces are easier to verify, e.g. using prover-verifier games (Kirchner et al., 2024), or rewarding reasoning which can be continued by weaker models (West et al., 2026).                                       |
| <b>Faithfulness training</b>                | Untrusted model itself                  | Training models so their reasoning traces accurately reflect their internal computations (Li et al., 2025).   |
| <b>CoT monitor training</b>                 | Untrusted model itself / Separate model | Training a model to review chains-of-thought or action traces and flag suspicious behaviour (Greenblatt, 2025c).  |
| <b>Activation oracles</b>                   | Untrusted model itself / Separate model | Supervised training of natural-language explainers of activations, on labelled (activation, question, answer) examples (Karvonen et al., 2025).   |
| <b>Natural-language autoencoders (NLAs)</b> | Untrusted model itself / Separate model | Unsupervised training of a paired activation-verbaliser (AV) and reconstructor, yielding free-text explanations of activations (Fraser-Taliente et al., 2026).  |
| <b>Sparse Autoencoders (SAEs)</b>           | Separate model                          | Training encoder-decoders to decompose activations into a dictionary of interpretable features (Cunningham et al., 2023; Bricken et al., 2023).   |
| <b>Supervised probes</b>                    | Separate model                          | Training supervised classifiers on internal activation states (Alain & Bengio, 2016; Belinkov, 2022).   |
| <b>Training against lie-detectors</b>       | Untrusted model itself                  | Attempting to increase honesty using white-box features (e.g. deception probe outputs) as a reward signal during training (though there is also a risk of obfuscating against the detector) (Cundy & Gleave, 2025).                                 |

**Constitutional AI and character training have been deployed in production, but they rely on broad generalisation.** Constitutional AI and character training (Askell et al., 2026; Maiya et al., 2025) shape a model’s default persona towards honesty and alignment. Both rely heavily on *broad generalisation*: a loosely specified set of principles or character descriptions is expected to translate across the full range of deployment. Rather than narrowly enhancing just oversight, these techniques are primarily intended to create a broadly aligned character. When these methods fail to provide alignment, failures to provide oversight will likely be correlated.

**Several training approaches aim to enhance or preserve monitorability via model outputs — through honesty, confession, verbalisation, or legibility.** As necessity-based arguments for CoT monitorability degrade, the ability to interrogate models and get honest answers becomes substantially more important; experts identify the broad area of honesty and confession training as the most natural candidates for this.<sup>61</sup> Confession and self-report training (Joglekar et al., 2025; Li et al., 2026; Lee et al., 2026) creates an incentive for models to self-report misbehaviour or hidden objectives, either when prompted or unprompted, using reward decoupling so that honest disclosure does not reduce the model’s task reward. Verbalisation propensity training (Turpin et al., 2025) trains models to acknowledge which factors influence their decisions. Legibility training (Kirchner et al., 2024; West et al., 2026) keeps reasoning traces verifiable by weaker models or humans, though at a measurable accuracy cost (a “legibility tax”). Although many experts are optimistic about this area of approaches, some note that auxiliary objectives have historically struggled to remain relevant in production-scale reinforcement learning.<sup>62</sup>

**Activation oracles, natural-language autoencoders, and introspection training are promising longer-term directions at the intersection of training and white-box methods.** Four experts singled out these directions as a particularly exciting longer-term direction (see subsection 5.1.2 for explanations), seeing them as potentially more robust than text-only methods because they access internal states directly, and so can optimise for CoT which is faithful to actual computations. However, current implementations remain difficult to use reliably (Jakkli et al., 2026; Fraser-Taliente et al., 2026), and may be most valuable as hypothesis-generation tools that narrow the search space for more expensive validation techniques.

**Some training-based approaches do not require training the untrusted model itself.** Trusted CoT monitor training, supervised probes, and unsupervised interpretability methods (e.g. SAEs) all train a separate model or classifier to oversee the untrusted model, avoiding some risks from the untrusted model being potentially misaligned. Their main limitation is that the classifiers themselves are weaker, and must rely on features (e.g. outputs or activations) provided by the untrusted model, which must also be labelled correctly for training.

## 6.2 Challenges for training interventions

| Oversight Degradation Pathways - Training for Oversight     |          |   |          |
|---|----------|---|----------|
| Degradation pathway   | Timeline | Likelihood                                  | Severity |
| <b>Adversarial generalisation failure (training-gaming)</b> | Ongoing  | Almost certain to become harder to rule out | ●        |
| <b>Training data representativeness</b>                     | Ongoing  | Highly likely to persist                    | ●        |
| <b>Non-adversarial generalisation failure</b>               | Ongoing  | Unclear; expert disagreement                | ●        |
| <b>Competing training objectives</b>                        | Ongoing  | Likely                                      | ●        |

<sup>61</sup>Of experts who commented on training-based fallbacks for CoT monitorability, around six endorsed honesty or confession training as a leading candidate; two explicitly framed it as the natural fallback to CoT, with one ranking it second in importance only to CoT itself.

<sup>62</sup>Of experts who commented on this area, four to five expressed broad optimism about honesty or confession training; one expert is sceptical, citing a poor track record for auxiliary objectives in production-scale RL, and one further expert is “*instinctively pessimistic*” about confessions specifically.

**Training interventions for improved oversight face a dual problem: (1) the generalisation problem, and (2) the training data problem.** The generalisation problem refers to a cluster of systematic reasons why training may fail to generalise to safety-critical deployment situations. The training-data problem refers to the difficulty of obtaining training data that is representative of the failures that oversight needs to catch.

**The training data problem is dual to the generalisation problem.** Insofar as generalisation is weaker, training data needs to be closer to the actual deployment situations in which oversight is required. Some approaches lean on broad generalisation (e.g. character training), while others aim to target their training data more narrowly to actual on-policy failures (e.g. confession training). Approaches relying on generalisation are more convenient as they require less data coverage, but their success is also more likely to be correlated with that of broader alignment training.

**Training interventions may fail because training doesn’t generalise from the available data to real deployment situations.** Experts raised two kinds of generalisation failure:

1. **Adversarial Generalisation Failures.** As a form of oversight awareness (Section 3.2), models able to distinguish training data from deployment may be able to engage in **Goal Guarding** – subverting training intended to make them monitorable.<sup>63</sup> For example, instrumentally behaving well during training may be a viable strategy to avoid being updated (Carlsmith, 2023; Hebbar & Greenblatt, 2025). This applies centrally to approaches involving training the untrusted model itself (e.g. confession, verbalisation, legibility, faithfulness, character, or untrusted CoT monitor training). Note that other forms of adversarial generalisation failures are also possible.<sup>64</sup> **As models become more capable of training subversion, it will almost certainly become more difficult to rule out adversarial generalisation failures.** This is closely related to evaluation awareness (Section 3.2): the same situational awareness that enables models to game evaluations or monitors could also enable them to identify and subvert training interventions aimed at improving oversight.
2. **Non-Adversarial Generalisation Failures.** Approaches may naturally fail to generalise across distributional shifts from training to deployment, or across coverage gaps in the training data – especially when these gaps are large and meaningful. The result would be *context-dependent monitorability* (analogous to *context-dependent misalignment* of MacDiarmid et al. (2025)).

#### Expert Perspectives

**Experts disagree sharply over the difficulty of the generalisation problem.** While three experts argued that adequate generalisation is the default outcome of training for oversight (e.g., honest character training or training a CoT monitor), two experts on the more pessimistic end maintain that sufficiently capable models will naturally separate training-like from deployment-like inputs, so that establishing effective oversight on training-like inputs

<sup>63</sup>Of experts who commented, one advanced this adversarial-generalisation framing most directly; three further experts treated it as a central reason for caution about training the untrusted model itself for oversight, while disagreeing on how decisive the concern is.

<sup>64</sup>Some types of reinforcement learning could potentially be subvertable by exploration hacking (Shlegeris & Stastny, 2025), which does not require distinguishing training from deployment. This merely requires monitors to sandbag (underperform) consistently during training. Additionally, even when the untrusted model is not directly being trained, it is still often involved in the process of generating training data. For example, supervised probes and trusted monitor training approaches may involve collecting training data from the untrusted model on a particular distribution. Speculatively, the untrusted model may be able to sabotage the training-processes of these methods if it has sufficient situational awareness and ability to cause the training distributions to be unrepresentative of real failures.

does not suffice to ensure adequate oversight in real deployment contexts. Moreover, these latter experts argue that this problem is insoluble without something close to a full alignment solution.

**It is hard to collect training data that is sufficiently representative of the failures that oversight needs to catch in practice.** The main reasons for this are:

1. **Circularity.** The most realistic and representative training data is real, on-policy examples of misaligned or deceptive behaviour. However, there’s a threat of circularity: collecting this data already presupposes having the tools to effectively monitor for real misalignment and deception, which is what the training is trying to achieve.<sup>65</sup> In practice, iterative deployment can partially mitigate this: once some examples are caught, training for monitorability of those examples could create a positive feedback loop, allowing more examples to be found. However, this also risks differentially increasing monitorability of the kinds of examples which are most easily caught, leaving gaps for novel failure modes.

2. **Coverage gaps & distributional shifts.** Three experts emphasised coverage gaps as a primary concern, highlighting that **we are unlikely to be able to get full-coverage on-distribution training data for the most concerning failures before they occur.** For example:

- Training data may cover known bad patterns of behaviour (e.g., egregious reward hacking) while missing novel or more subtle failures (e.g., research sabotage).
- It may be challenging to elicit examples of the untrusted model itself being misaligned (on-policy), so synthetic transcripts elicited from another model or written by humans (off-policy) may be used as training data. Models may be able to distinguish misaligned transcripts that they have produced themselves from off-policy training data, potentially allowing them to subvert training.

3. **Reward specification issues.** Even when training with on-policy reinforcement learning, the reward signal for “monitorability” is often a proxy for the property actually desired. For example, confession training rewards text that a judge rates as an honest disclosure, not honest disclosure itself (Joglekar et al., 2025). The model may learn to satisfy the proxy without achieving the intended property — such as by producing plausible-sounding confessions that omit the real issue.<sup>66</sup>

**Monitorability sometimes stands opposed to other training objectives.** Training for oversight interacts with the rest of the training pipeline in ways that can undermine monitorability.

- **Capability tension.** Verbalisation and legibility objectives may directly compete with capability. If verbalising reasoning is useful for the task, models do it naturally; if not, the monitorability objective fights the capability objective. Legibility training via prover-verifier games incurs a measurable accuracy cost — a “legibility tax” — relative to optimising for correctness alone (Kirchner et al., 2024).<sup>67</sup>

<sup>65</sup>Two experts named this circularity directly: one observed that “confessions work best when... your model is scheming, you’ve caught it doing egregiously bad stuff, and you train it to reveal those things”, and another noted that for untrusted-monitor training “I don’t see where you get the training data”.

<sup>66</sup>One expert flagged this proxy gap directly, noting that confession training “ends up bottlenecked by your ability to inspect whether the confession is actually reliable”, so labs effectively distil their best guess of honesty rather than honesty itself. A second expert raised related concerns about the difficulty of establishing ground truth for honesty more broadly.

<sup>67</sup>One expert highlighted this competition explicitly: “either being legible in the chain of thought was actually useful for your task and the model does it, or it’s not”.

- **Inference cost pressure.** There is commercial pressure to reduce inference costs. Training to reduce token usage undermines monitorability by shrinking the surface area available for CoT monitoring, and may push reasoning into latent computation that is harder to oversee (see [section 2](#)).
- **Unpredictable interactions with subsequent training.** More broadly it is difficult to predict how monitorability properties will change under training processes not directly aimed at affecting them. For example, training to avoid revealing sensitive information in prompt-injection attacks might generalise unexpectedly to undermining the model’s forthcomingness in other domains.<sup>68</sup>

### 6.3 Developing and stress-testing training interventions

**Continue to develop training-based approaches for oversight.** Most training-based approaches for oversight are new and under-explored, and experts identified this as a key area with significant room for improvement. Confession and honesty training, verbalisation propensity training, legibility training, and faithfulness training are all early-stage research directions that could become substantially more effective with further investment. Improving these techniques in realistic settings is particularly important given that they may serve as fallbacks if other monitoring avenues (e.g. CoT monitoring) degrade.

**Stress-testing training interventions.** There are stress-testing frameworks designed to probe the susceptibility of training interventions to various failure modes:

1. **Malign initialisation experiments.** These experiments are designed to test the robustness of training against malign initialisations – misaligned models specifically designed to persist through oversight training.
2. **Robustness to compromised training processes.** These experiments are designed to test whether oversight training processes are robust to being partially compromised. For example, is the oversight-enhancing training sufficient even when other reward signals favour deception?
3. **Realistic failed training process.** These experiments try to discover the most realistic training processes which result in misalignment or diminished oversight.

---

<sup>68</sup>One expert raised this directly: “I don’t think we’re able to make good predictions about any of those forms of honesty and how they’re going to change with some change to a training process that’s not directly pointed at affecting that”.

## 7 Summary and Recommendations

*The following discussion represents the views of the project team at UK AISI, informed by but not necessarily representative of all experts interviewed. Readers are encouraged to consider the limitations of our methods and this report discussed in [Appendix B](#).*

### 7.1 Summary

**Current oversight rests on properties which are likely to erode over the next few years, absent intervention to preserve them.** The preceding sections have examined five oversight domains and identified more than twenty distinct degradation pathways. Some of these pressures are already visible: single-forward-pass capabilities are improving ([subsubsection 2.2.2](#)), evaluation awareness is affecting alignment audits ([subsection 3.2](#)), and opportunities for low-incrimination harmful actions are increasing as deployment settings become more complex and long-horizon ([subsection 3.1](#)). Others are not yet present but risk severely undermining current monitoring approaches, such as architecture shifts to latent reasoning ([subsubsection 2.2.1](#)), training-gaming ([subsection 6.2](#)), representational drift ([subsection 5.2](#)), or opaque memory and communication systems ([section 4](#)). While many individual pathways are uncertain, the breadth and variety of pressures on oversight is cause for concern.

**These degradation pathways are likely to interact in ways that compound risk — such as through training feedback loops.** Concerningly, even a small initial reduction in monitorability (e.g. from linguistic drift or no-CoT capabilities) could be amplified by training processes that reward undetected misbehaviour. If a model discovers that it can reward-hack without being caught in certain contexts, then this behaviour gets reinforced, driving further degradation. Other interactions matter too: loss of chain-of-thought legibility makes white-box signals harder to ground ([subsection 5.2](#)); and greater coherence, opaque communication and memory, and situational awareness may together produce systems more capable of stealth, collusion, or selective sandbagging than any one of these changes would suggest in isolation.

**Some loss of monitorability could go unnoticed.** Many oversight-relevant properties are not currently tracked in a way that would reliably reveal their degradation. There is plenty of room for better measurement across oversight domains, including white-box oversight ([subsection 5.3](#)), training generalisation ([section 6](#)), and the integrity of evaluations ([subsection 3.2](#)).

**Finally, the techniques discussed in this report are largely aimed at the regime before full automation of AI R&D; they are not a plan for overseeing substantially superhuman systems.** A sufficiently capable model could in principle defeat every oversight surface we consider here — exhibit fully legible CoT while reasoning in latent space, pass every alignment honeypot while behaving badly in deployment, produce internal activations that confound probes and NLAs, and route coordination through channels that no single human can audit. Our claim is not that the levers below solve this problem, but that they buy time and evidence in the intermediate regime, and can be marginally useful for lowering risks. Safety planning for substantially superhuman systems requires additional ingredients beyond the scope of this report.

### 7.2 The challenge of acting on oversight evidence

Translating evidence from oversight techniques into action is harder than it might appear.

**The evidence derived from monitoring and auditing often rests on subtle, complicated, or contested technical assumptions.** For example, probes on an AI model’s internal activations ([subsubsection 5.1.1](#)) produce numerical values that are difficult to interpret and

trust without a large amount of context on the probe training and validation process.<sup>69</sup> This means that even when oversight methods are accurate, they can fail to convince decision-makers with different prior beliefs or foundational assumptions. The most useful oversight techniques are those which are not only accurate, but produce legible, easily verifiable evidence that is robust to these differences in assumptions.

**Requiring strong, legible evidence before taking action risks biasing decisions towards inaction.** Experts largely agreed that the most compelling evidence is clear, demonstrable cases where the model has taken an egregiously misaligned action in an actual deployment-like setting. However, by the time such evidence is available, it may be too late to intervene effectively. There is a risk that requiring clear, demonstrable cases of egregiously misaligned behaviour before acting sets too high a bar, missing key windows of opportunity for effective safety interventions. Safe development and deployment of advanced AI may require taking action even on inconclusive or ambiguous evidence: this should be handled carefully, by deciding in advance where the burden of proof lies and what would count as sufficient evidence to conservatively limit risks, watching closely for early signals, and developing well-reasoned methods to perturb systems and stress-test for signs of misalignment or vulnerability to misuse.

**Acting on oversight evidence could mean making expensive decisions, but could also meaningfully reduce risks.** Halting a deployment, retraining, or rolling back to an earlier checkpoint has direct, attributable costs in lost revenue, competitive speed, or foregone research, whereas the cost of *not* acting on a concerning signal is typically diffuse and externalised.<sup>70</sup> The benefits of mitigating risks which do not eventuate, plausibly due to critical decisions, are inherently less salient. This asymmetry biases decisions towards inaction. One potential intervention is reducing the costs of acting (e.g. through coordination or technical research); another is internalising the costs from risks which would otherwise be felt externally, incentivising model developers to be more conservative.

### 7.3 Recommendations

Here we highlight our judgements of the most important recommendations, building on the findings discussed throughout the report. We organise these into four areas: implementing currently available oversight techniques; measuring exposure to oversight degradation; preserving current oversight channels; and advancing emerging techniques that may be required if currently available channels continue to degrade.

#### 7.3.1 Implementing currently available oversight

**AI developers and deployers should implement and maintain a diverse set of oversight techniques, providing defence in depth against the failure of any single surface.** Some past failures would likely have been preventable by implementing existing oversight techniques such as well-targeted evaluation and monitoring (OpenAI, 2025b,c), rather than requiring entirely new methods. Large benefits can still be gained by applying existing oversight techniques while they continue to work well, including CoT monitoring (subsubsection 2.1.4), action-only monitoring and AI control (subsection 3.1; Greenblatt et al., 2024b; Bhatt et al., 2025), evaluations (subsection 3.2), training (section 6), and white-box methods such as probes (subsubsection 5.1.1).

---

<sup>69</sup>Probing an AI model’s internal activations to monitor for deception requires that human-analogous deception representations are used by the model, and that the probe training process has targeted these concepts sufficiently well. Both of these pieces are hard to demonstrate conclusively, let alone concisely, to a non-technical decision-maker.

<sup>70</sup>This may be compounded by fragmented authority across the supply chain (original model developer, scaffolding developer or fine-tuner, API deployer, end user), with no single party having sole responsibility to implement monitoring or to act on a concerning signal.

At minimum, this means:

- running regular evaluations or beta-tests with broad coverage of relevant risks;
- restricting the damage a model could do, e.g. via sandboxing and synchronous monitoring;
- retaining logs sufficient to reproduce and investigate concerning behaviours, and asynchronously monitoring those logs for concerning patterns.<sup>71</sup>

However as oversight-relevant properties degrade and models become more capable, the techniques that need to be implemented will themselves need to advance. The research priorities below are aimed at ensuring that effective oversight techniques continue to exist.

### 7.3.2 Measuring exposure to oversight degradation

**There are strong reasons for developers to understand their exposure to oversight degradation.** Oversight is a systemic factor underpinning safety, and changes to it can be gradual and hard to notice. Developers should map which of their safety claims depend on which oversight properties, so that when a property degrades, the downstream consequences are understood. Risk assessments should include forward projections of whether key oversight properties will hold (see [subsection 2.3](#) and [subsection 5.3](#)), so that safety arguments do not silently expire. Some risk assessments already include basic forward-looking discussion of this kind ([Anthropic, 2026](#)), which could be expanded and become standard practice. We highlight three specific measurement priorities:

1. **Further develop monitorability benchmark suites.** Benchmarks measuring how monitorable frontier models are should be implemented and run to compare models and trends over time across monitoring approaches (action-only, CoT, white-box), task types (code, research, decision-making), and risk models (hacking, sandbagging, research sabotage). Early work in this area ([Kutasov et al., 2025a](#); [Guan et al., 2025](#); [Emmons et al., 2025b](#); [Kinniment et al., 2026](#); [Chen et al., 2026](#)) provides a foundation, but coverage remains incomplete in risk-relevant domains, and it is difficult to make comparisons between different model providers. Such a benchmark could become a standard component of model evaluations and system cards.
2. **Investigate whether representational drift across training undermines white-box tools.** It is not yet clear whether representational drift during RL training will become a serious practical problem for white-box oversight ([subsubsection 5.2.1](#)). Some evidence suggests that RL can degrade even freshly trained probes ([Taufeeque et al., 2026](#)), but the evidence is mixed. Model developers should carefully measure how well probes and SAEs trained on earlier checkpoints transfer to later ones, and how well newly trained probes perform, to determine whether this concern is arising.
3. **Distinguish different dimensions of CoT monitorability.** As well as measuring monitorability as a whole, we believe it is also valuable to distinguish whether a model verbalises because it *needs* to ([subsubsection 2.1.1](#)) or merely has a *propensity* to ([subsubsection 2.1.2](#)). Oversight based on propensity is likely much more fragile against training pressure or misalignment, so tracking them separately is important for understanding and mitigating risks. Although separating capabilities from propensities can be difficult, there are techniques such as running no-CoT evaluations ([Greenblatt, 2025b](#)), or ensuring that capabilities for oversight subversion are fully elicited in capabilities evaluations. Finally, human-legibility should also be tracked, separately from machine-monitorability, as legibility may degrade even if monitorability remains ([subsubsection 2.1.3](#)).

---

<sup>71</sup>Including during training.

### 7.3.3 Preserving current oversight

**Several of the highest-impact research directions are about preventing degradation of oversight properties we already have, rather than building new capabilities from scratch.** There is a window of opportunity now to study and preserve current oversight channels while they remain strong.

- 4. Report deliberate decisions that affect oversight properties.** Developers should report on oversight-relevant architecture changes (such as adoption of latent reasoning architectures (subsection 2.2.1) or continual learning mechanisms (subsection 4.2)) and training processes (such as training against monitors – section 2.2.3), and measure their exposure to shifts in oversight-relevant properties. Safety teams should be given sufficient notice ahead of time to thoroughly assess the downstream impacts of these changes.
- 5. Do not implement changes without somehow maintaining oversight.** When a development decision would degrade an oversight property (e.g., compressing chain-of-thought (section 2.2.3), adopting vector-based memory (subsection 4.2)), developers should demonstrate that alternative oversight mechanisms are in place to maintain monitorability before proceeding, and that they are sufficiently robust.
- 6. Map which training signals degrade or improve monitorability.** Different forms of optimisation pressure – outcome-based RL, length penalties, training against monitors, process reward models, deliberative alignment – likely have very different effects on CoT monitorability and internal representations. Empirically characterising these effects would provide concrete guidance on which training choices to avoid or prefer, and would help quantify the “monitorability tax” of different approaches. Baker et al. (2025) and Kaufmann et al. (2026) provide early work on this. Developers should also establish norms to disclose instances where they have unintentionally trained against monitorability, as has already occurred in practice (Bowman et al., 2025; METR, 2025; Anthropic, 2026).
- 7. Measure the correlation of monitoring surfaces.** Our central recommendation is to hold out decorrelated monitoring surfaces from training (subsection 2.4), but we do not yet know which surfaces are correlated with which, or in what ways. More empirical work is needed to map these relationships. For example: if a model is trained against a CoT monitor, does probe accuracy on the same tasks also degrade? Do action-only monitors, CoT monitors, and probes tend to catch the same misbehaviour instances, or different ones? Answering these questions would tell developers which surfaces actually provide independent evidence and which give correlated false confidence, making our hold-out recommendation concrete enough to act on.

### 7.3.4 Advancing emerging oversight techniques

**If current oversight channels continue to degrade, the field will need to rely increasingly on techniques that are currently immature.** Investing in these techniques now, while they can still be validated against current strong oversight signals, is essential.

- 8. Build model organisms of the oversight degradation pathways discussed in this report.** Model organisms are AI systems constructed to exhibit a target behaviour or property. Building realistic models exhibiting the major oversight degradation pathways would let oversight techniques be developed and stress-tested against the relevant failures before they emerge in frontier models.
- 9. Increase the ability to diagnose misalignment from suspicious incidents.** Establishing with confidence that a model is in fact misaligned (rather than confused, overeager, or jailbroken) is currently difficult. Better forensic infrastructure such as comprehensive

logging that enables replay of incidents, the ability to re-run incidents with perturbations, and investigator agents that can run many counterfactual rollouts using interpretability tools should be a major priority of oversight research.

10. **Stress-test training-based oversight against realistic failure modes.** Approaches such as character training, confession training, and legibility training (section 6) are promising but largely untested in adversarial conditions. Key experiments include: (a) malign initialisations (subsection 6.2), testing whether various training approaches can reliably override pre-existing strategies designed to subvert them; (b) testing generalisation boundaries, to characterise where training-based oversight works and where it fails; and (c) detecting context-dependent misalignment (subsubsection A.3.2), where safety training fixes problems only near the training distribution while leaving them present in untested situations.
11. **Advance white-box auditing towards production readiness.** Probes, SAEs, activation oracles, and natural-language autoencoders (section 5) are promising techniques that access internal states directly, making them potentially more robust against scheming-related suppression of text-based reasoning. However, current implementations remain difficult to use reliably on safety-relevant tasks (subsection 5.2). Priority work includes: improving the reliability and usefulness of activation oracles and natural-language autoencoders (and red-teaming them e.g. with malign initialisations); developing interpretability tools that scale to long reasoning traces and complex agentic settings; and reducing practical barriers to integrating probes into production inference stacks (subsubsection 5.2.7). Deploying probes for the most effective use cases now (e.g. CSAM detection, cyber-misuse) can establish the infrastructure needed for more ambitious safety probes later. Investment in open-source replications of white-box techniques would also help proliferate these tools across the ecosystem.
12. **Advance the science of production-based evaluations.** Production-based evaluations (replaying real deployment situations with untrusted models, OpenAI, 2025a) are among the most promising mitigations against evaluation gaming, as models exhibit substantially lower evaluation awareness on real deployment data (subsubsection 3.2.4). However, extending this approach to long trajectories, novel risk scenarios, and complex scaffolding remains an open challenge.
13. **Develop speculative mitigations to scenario-driven evaluation gaming.** Scenario-driven evaluation awareness – where models recognise the overall situation as an evaluation (subsubsection 3.2.2) – is likely to become the dominant challenge if micro-detail-driven awareness is addressed through more realistic evaluation design. Activation steering to suppress evaluation-awareness representations (Nguyen et al., 2025; Hua et al., 2025) is a promising but immature approach. Research should characterise where steering works and fails, and whether sufficiently capable models can resist or circumvent it.
14. **Prepare monitoring tools for architecture transitions.** If latent reasoning architectures (subsubsection 2.2.1), vector-based memory, or non-text inter-agent communication (section 4) are adopted, current monitoring approaches will be severely undermined. Proactive investment in interpretability tools for these modalities – before they are adopted at scale – would avoid a situation where monitoring capabilities lag far behind deployment realities. This includes adapting existing activation-level techniques (subsection 5.4) to work on non-text internal representations, and developing tools for inspecting and editing vector-based memory stores.

## Acknowledgements

We are grateful for the experts listed in Figure 1 for their insights during the interviews, and substantive feedback on this document. Additionally, thanks to Sid Black, Satvik Golechha, Nix Goldowsky-Dill, Charlie Griffin, Rob Kirk, Kola Ayonrinde, Hadrien Pouget, Alan Cooney, Merlin Stein, Caitlin van Bommel, Aliya Ahmad, James Burn, Nate Burnikell and Ture Hinrichsen for valuable review and discussions.

## References

- Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.04697>.
- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL <https://arxiv.org/abs/1610.01644>.
- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>.
- Anthropic. Eval awareness in Claude Opus 4.6’s BrowseComp performance, 2025a. URL <https://www.anthropic.com/engineering/eval-awareness-browsecomp>.
- Anthropic. Claude Opus 4.6 System Card, 2025b. URL <https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf>.
- Anthropic. Claude Opus 4 & Claude Sonnet 4 System Card, 2025c. URL <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>.
- Anthropic. Claude Sonnet 4.5 System Card, 2025d. URL <https://www.anthropic.com/claude-sonnet-4-5-system-card>.
- Anthropic. Alignment Risk Update: Claude Mythos Preview, 2026. URL <https://www-cdn.anthropic.com/79c2d46d997783b9d2fb3241de43218158e5f25c.pdf>.
- Anthropic. System Card: Claude Mythos Preview. <https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf>, 2026.
- Anthropic. Sabotage Risk Report: Claude Opus 4.6, 2026. URL <https://www-cdn.anthropic.com/f21d93f21602ead5cdbecb8c8e1c765759d9e232.pdf>.
- Santiago Aranguri and Tom McGrath. Discovering undesired rare behaviors via model diff amplification. *Goodfire*, 2025. URL <https://www.goodfire.ai/research/model-diff-amplification>.
- Rauno Arike. On Recent Results in LLM Latent Reasoning, 2025. URL <https://www.lesswrong.com/posts/pLnLSgWphqDbdorgi/on-recent-results-in-llm-latent-reasoning>.
- Amanda Askell, Joe Carlsmith, and Anthropic. Claude’s Constitution, 2026. URL <https://www.anthropic.com/constitution>.

- Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated Activations Bypass LLM Latent-Space Defenses. *arXiv preprint arXiv:2412.09565*, 2024. URL <https://arxiv.org/abs/2412.09565>.
- 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, 2025. URL <https://arxiv.org/abs/2503.11926>.
- Mikita Balesni, Tomek Korbak, and Owain Evans. Lessons from Studying Two-Hop Latent Reasoning, 2025. v4, revised Nov 2025.
- Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. Toward Universal Steering and Monitoring of AI Models. *arXiv preprint arXiv:2502.03708*, 2025. URL <https://arxiv.org/abs/2502.03708>.
- Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to Memorize at Test Time. *arXiv preprint arXiv:2501.00663*, 2025. URL <https://arxiv.org/abs/2501.00663>.
- Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. URL <https://arxiv.org/abs/2102.12452>.
- Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger Grosse, Samuel R. Bowman, and David Duvenaud. Sabotage Evaluations for Frontier Models. *arXiv preprint arXiv:2410.21514*, 2024. URL <https://arxiv.org/abs/2410.21514>.
- Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-Z: Controlling AI Agents via Resampling. *arXiv preprint arXiv:2504.10374*, 2025. URL <https://arxiv.org/abs/2504.10374>.
- Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting Neural Networks through the Polytope Lens. *arXiv preprint arXiv:2211.12312*, 2022. URL <https://arxiv.org/abs/2211.12312>.
- Nelson Jonathan Bostock, Georgiy Kozhevnikov, Sinclair Gardner-Challis, Joan Abate, Alessandro Griffin, and Charlie Velja. When Can We Trust Untrusted Monitoring? A Safety Case Sketch Across Collusion Strategies. *arXiv preprint arXiv:2602.20628*, 2026. URL <https://arxiv.org/abs/2602.20628>.
- Samuel R. Bowman, Misha Wagner, Fabien Roger, and Holden Karnofsky. Pilot Sabotage Risk Report. Anthropic, 2025. URL <https://www.anthropic.com/research/sabotage-risk-report-summer-2025>.
- Gerard Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. *arXiv preprint arXiv:2509.21344*, 2025. URL <https://arxiv.org/abs/2509.21344>. Originally titled “Towards mitigating information leakage when evaluating safety monitors”.
- Dan Braun, Lucius Bushnaq, Stefan Heimersheim, Jake Mendel, and Lee Sharkey. Interpretability in Parameter Space: Attribution-based Parameter Decomposition. *arXiv preprint arXiv:2501.14926*, 2025. URL <https://arxiv.org/abs/2501.14926>.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan

- Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, 2023. URL <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Joe Carlsmith. The Goal-Guarding Hypothesis (Section 2.3.1.1 of “Scheming AIs”), December 2023. URL <https://www.alignmentforum.org/posts/LJpqnbgaoTJL6iMAf/the-goal-guarding-hypothesis-section-2-3-1-1-of-scheming-ais>. Alignment Forum.
- Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, J r my Scheurer, Marius Hobbhahn, et al. Black-Box Access is Insufficient for Rigorous AI Audits, 2024. URL <https://arxiv.org/abs/2401.14446>.
- Maheep Chaudhary and Atticus Geiger. Evaluating Open-Source Sparse Autoencoders on Disentangling Factual Knowledge. *arXiv preprint arXiv:2409.04478*, 2024. URL <https://arxiv.org/abs/2409.04478>.
- Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning, 2025a. URL <https://arxiv.org/abs/2505.16782>.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don’t Always Say What They Think, 2025b. URL <https://arxiv.org/abs/2505.05410>.
- Yueh-Han Chen, Robert McCarthy, Bruce W. Lee, He He, Ian Kivlichan, Bowen Baker, Micah Carroll, and Tomek Korbak. Reasoning Models Struggle to Control Their Chains of Thought, 2026. URL <https://arxiv.org/abs/2603.05706>.
- James Chua and Owain Evans. Are DeepSeek R1 And Other Reasoning Models More Faithful? *arXiv preprint arXiv:2501.08156*, 2025. URL <https://arxiv.org/abs/2501.08156>.
- Alex Cloud, Jacob Goldman-Wetzler, Even Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient Routing: Masking Gradients to Localize Computation in Neural Networks, 2024. URL <https://arxiv.org/abs/2410.04332>.
- Russell Coleman. Eval awareness in Claude Opus 4.6’s BrowseComp performance. Anthropic Engineering Blog, 2026. URL <https://www.anthropic.com/engineering/eval-awareness-browsecomp>.
- Asa Cooper Stickland, Jan Michelfeit, Arathi Mani, Charlie Griffin, Ollie Matthews, Tomek Korbak, Rogan Inglis, Oliver Makins, and Alan Cooney. Async Control: Stress-testing Asynchronous Control Measures for LLM Agents. *arXiv preprint arXiv:2512.13526*, 2025. URL <https://arxiv.org/abs/2512.13526>.
- Chris Cundy and Adam Gleave. Preference Learning with Lie Detectors can Induce Honesty or Evasion. *arXiv preprint arXiv:2505.13787*, 2025. URL <https://arxiv.org/abs/2505.13787>.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint arXiv:2309.08600*, 2023. URL <https://arxiv.org/abs/2309.08600>.
- Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, Alek Dimitriev, Rob Gilson, Logan

- Howard, Yijin Hua, Jared Kaplan, Jan Leike, Mu Lin, Christopher Liu, Vladimir Mikulik, Rohit Mittapalli, Clare O’Hara, Jin Pan, Nikhil Saxena, Alex Silverstein, Yue Song, Xunjie Yu, Giulio Zhou, Ethan Perez, and Mrinank Sharma. Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks. *arXiv preprint arXiv:2601.04603*, 2025. URL <https://arxiv.org/abs/2601.04603>.
- Cursor AI. Real-Time RL for Composer. <https://cursor.com/blog/real-time-rl-for-composer>, 2025a. Accessed 2026-04-09.
- Cursor AI. Real-Time RL for Tab. <https://cursor.com/blog/tab-rl>, 2025b. Accessed 2026-04-09.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step, 2024. URL <https://arxiv.org/abs/2405.14838>.
- Jacob Drori, Luke Marks, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Output Supervision Can Obfuscate the Chain of Thought, 2025. URL <https://arxiv.org/abs/2511.11584>.
- Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders Find Interpretable LLM Feature Circuits, 2024. URL <https://arxiv.org/abs/2406.11944>.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- Scott Emmons, Erik Jenner, David K. Elson, Rif A. Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors, 2025a. URL <https://arxiv.org/abs/2507.05246>.
- Scott Emmons, Roland S. Zimmermann, David K. Elson, and Rohin Shah. A Pragmatic Way to Measure Chain-of-Thought Monitorability, 2025b.
- Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear, 2024. URL <https://arxiv.org/abs/2405.14860>.
- Ching Fang and Samuel Marks. Unsupervised Decoding of Encoded Reasoning Using Language Model Interpretability. *arXiv preprint arXiv:2512.01222*, 2025. URL <https://arxiv.org/abs/2512.01222>.
- Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarnig, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations. *Transformer Circuits Thread*, May 2026. URL <https://transformer-circuits.pub/2026/nla/index.html>.
- Leo Gao. An Ambitious Vision for Interpretability, 2025. URL <https://www.alignmentforum.org/posts/Hy6PX43HGgmfiTaKu/an-ambitious-vision-for-interpretability>.
- Johannes Gasteiger, Abir Khan, Sam Bowman, Misha Wagner, Vladimir Mikulik, Ethan Perez, and Fabien Roger. Automated Researchers Can Subtly Sandbag. 2025. URL <https://alignment.anthropic.com/2025/automated-researchers-sandbag/>.

- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. *arXiv preprint arXiv:2502.05171*, 2025. doi: 10.48550/arXiv.2502.05171. URL <https://arxiv.org/abs/2502.05171>.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting Strategic Deception Using Linear Probes. *arXiv preprint arXiv:2502.03407*, 2025. URL <https://arxiv.org/abs/2502.03407>.
- Google DeepMind. Gemini Diffusion. <https://deepmind.google/models/gemini-diffusion/>, 2025. Accessed 2026-04-09.
- Liv Gorton. What Would Non-Linear Features Actually Look Like? <https://livgorton.com/non-linear-feature-reps>, 2026. Blog post, January 27, 2026.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Ryan Greenblatt. Recent LLMs can use filler tokens or problem repeats to improve (no-CoT) math performance, 2025a. URL <https://blog.redwoodresearch.org/p/recent-llms-can-use-filler-tokens-or-problem-repeats-to>. Redwood Research Blog, December 2025.
- Ryan Greenblatt. Measuring no CoT math time horizon, 2025b. URL <https://blog.redwoodresearch.org/p/measuring-no-cot-math-time-horizon>. Redwood Research Blog.
- Ryan Greenblatt. Prefix Cache Untrusted Monitors: A Method to Apply After You Catch Your AI, 2025c. URL <https://blog.redwoodresearch.org/p/prefix-cache-untrusted-monitors-a>. Redwood Research blog.
- Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment Faking in Large Language Models. *arXiv preprint arXiv:2412.14093*, 2024a. URL <https://www.anthropic.com/research/alignment-faking>.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI Control: Improving Safety Despite Intentional Subversion. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b. URL <https://proceedings.mlr.press/v235/greenblatt24a.html>.
- Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative Alignment: Reasoning Enables Safer Language Models, 2024. URL <https://arxiv.org/abs/2412.16339>.
- Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring Monitorability, 2025. URL <https://arxiv.org/abs/2512.18311>.
- Rohan Gupta and Erik Jenner. RL-obfuscation: Can language models learn to evade latent-space monitors? *arXiv preprint arXiv:2506.14261*, 2025. URL <https://arxiv.org/abs/2506.14261>.

- Yoav Gur-Arieh, Clara Suslik, Yihuai Hong, Fazl Barez, and Mor Geva. Precise In-Parameter Concept Erasure in Large Language Models, 2025. URL <https://arxiv.org/abs/2505.22586>.
- Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777), 2025. doi: 10.1126/science.aea3884. URL <https://doi.org/10.1126/science.aea3884>.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. *arXiv preprint arXiv:2412.06769*, 2024. URL <https://arxiv.org/abs/2412.06769>.
- Vivek Hebbar. How can we solve diffuse threats like research sabotage with AI control?, 2025. URL <https://blog.redwoodresearch.org/p/how-can-we-solve-diffuse-threats>.
- Vivek Hebbar and Ryan Greenblatt. When Does Training a Model Change Its Goals?, June 2025. URL <https://www.alignmentforum.org/posts/yvuXPi5m4vCvSGTjo/when-does-training-a-model-change-its-goals>. Alignment Forum.
- Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. *NeurIPS*, 2025. URL <https://arxiv.org/abs/2503.01822>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss Landscape Degeneracy and Stagewise Development in Transformers, 2025. URL <https://arxiv.org/abs/2402.02364>.
- Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. ThinkPrune: Pruning long chain-of-thought of LLMs via reinforcement learning, 2025. URL <https://arxiv.org/abs/2504.01296>.
- Nikolaus Howe and Micah Carroll. The Ends Justify the Thoughts: RL-Induced Motivated Reasoning in LLM CoTs, 2025. URL <https://arxiv.org/abs/2510.17057>.
- Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering Evaluation-Aware Language Models to Act Like They Are Deployed. *arXiv preprint arXiv:2510.20487*, 2025. URL <https://arxiv.org/abs/2510.20487>.
- Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL <https://aclanthology.org/2024.acl-long.470/>.
- Evan Hubinger. Model Organisms of Misalignment: The Case for a New Pillar of Alignment Research. Alignment Forum, 2023. URL <https://www.alignmentforum.org/posts/ChDH335ckdvpXaXX/model-organisms-of-misalignment-the-case-for-a-new-pillar-of-1>.
- John Hughes, Alex Isken, Alexander Glynn, Conner Phillippi, David Dworken, Emily To, Fabien Roger, Jake Eaton, Javier Rando, Shawn Moore, and Soyary Sunthorn. Claude Code auto mode: a safer way to skip permissions, 2026. URL <https://www.anthropic.com/engineering/claude-code-auto-mode>.

- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis. *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL <https://arxiv.org/abs/2405.07987>.
- Rogan Inglis, Ollie Matthews, Tyler Tracy, Oliver Makins, Tom Catling, and Asa Cooper Stickland. ControlArena: A collection of settings, model organisms and protocols for running control experiments. UK AI Safety Institute, 2025. URL <https://github.com/UKGovernmentBEIS/control-arena>.
- Arya Jakkli, Senthoran Rajamanoharan, and Neel Nanda. Current Activation Oracles Are Hard to Use. LessWrong, 2026. URL <https://www.lesswrong.com/posts/LXQBcztrWKhtcgQfJ/current-activation-oracles-are-hard-to-use>. Research conducted during MATS 9.0.
- Manas Joglekar, Jeremy Chen, Gabriel Wu, Jason Yosinski, Jasmine Wang, Boaz Barak, and Amelia Glaese. Training LLMs for Honesty via Confessions, 2025. URL <https://arxiv.org/abs/2512.08093>.
- David Johnston and Nora Belrose. Examining Two Hop Reasoning Through Information Content Scaling, 2025.
- David O. Johnston, Arkajyoti Chakraborty, and Nora Belrose. Mechanistic Anomaly Detection for "Quirky" Language Models, 2025. URL <https://arxiv.org/abs/2504.08812>.
- Arun Jose. Reasoning Models Sometimes Output Illegible Chains of Thought, 2025. URL <https://arxiv.org/abs/2510.27338>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Adam Karvonen and Sam Marks. Race and Gender Bias As An Example of Unfaithful Chain of Thought in the Wild, July 2025. URL <https://www.lesswrong.com/posts/me7wFrkEtMbkzXGJt/race-and-gender-bias-as-an-example-of-unfaithful-chain-of>.
- Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers. *arXiv preprint arXiv:2512.15674*, 2025. URL <https://arxiv.org/abs/2512.15674>.
- Max Kaufmann, David Lindner, Roland S. Zimmermann, and Rohin Shah. Aligned, Orthogonal or In-conflict: When can we safely optimize Chain-of-Thought? 2026.
- Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Janis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. On Scalable Oversight with Weak LLMs Judging Strong LLMs. *arXiv preprint arXiv:2407.04622*, 2024. URL <https://arxiv.org/abs/2407.04622>.
- Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, and Volodymyr Kuleshov. Mercury: Ultra-Fast Language Models Based on Diffusion. *arXiv preprint arXiv:2506.17298*, 2025. URL <https://arxiv.org/abs/2506.17298>.

- Megan Kinniment, Seraphina Nix, Thomas Broadley, and Neev Parikh. Early work on monitorability evaluations, 2026. URL <https://metr.org/blog/2026-01-19-early-work-on-monitorability-evaluations/>. METR Blog, January 2026.
- Nathalie Kirch, Constantin Weisser, Severin Field, Helen Yannakoudakis, and Stephen Casper. What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks. *arXiv preprint arXiv:2411.03343*, 2024. URL <https://arxiv.org/abs/2411.03343>.
- Nathalie Kirch, Samuel Dower, Adrians Skapars, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. The Impact of Off-Policy Training Data on Probe Generalisation. *arXiv preprint arXiv:2511.17408*, 2025. URL <https://arxiv.org/abs/2511.17408>.
- Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-Verifier Games improve legibility of LLM outputs, 2024. URL <https://arxiv.org/abs/2407.13692>.
- Robert Kirk, Alexandra Souly, Kai Fronsdal, Abby D’Cruz, and Xander Davies. Evaluating whether AI models would sabotage AI safety research. UK AISI Blog, 2026. URL <https://www.aisi.gov.uk/blog/evaluating-whether-ai-models-would-sabotage-ai-safety-research>. Also arXiv preprint arXiv:2604.24618, submitted 2026-04-27.
- Connor Kissane, Monte MacDiarmid, and Fabien Roger. Measuring and Improving Coding Audit Realism with Deployment Resources. Anthropic Alignment Science Blog, March 2026. URL <https://alignment.anthropic.com/2026/coding-audit-realism/>.
- Daniel Kokotajlo and Abram Demski. Why Don’t We Just... Shoggoth+Face+Paraphraser?, November 2024. URL <https://www.lesswrong.com/posts/Tzdwetw55JNqFTkzK/why-don-t-we-just-shoggoth-face-paraphraser>.
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mdry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, 2025. URL <https://arxiv.org/abs/2507.11473>.
- János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building Production-Ready Probes For Gemini, 2026. URL <https://arxiv.org/abs/2601.11516>.
- Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars’ Bench: Evaluating Lie Detectors for Language Models, 2026. URL <https://arxiv.org/abs/2511.16035>.
- Gerson Kroiz, Greg Kocher, and Tim Hua. Can Models Be Evaluation Aware Without Explicit Verbalization?, 2025. URL <https://www.lesswrong.com/posts/W6ZFnheeEBGcZqdHd/can-models-be-evaluation-aware-without-explicit>.
- Kalle Kujanpää, Pekka Marttinen, Harri Valpola, and Alexander Ilin. Efficient Knowledge Injection in LLMs via Self-Distillation. *arXiv preprint arXiv:2412.14964*, 2024. URL <https://arxiv.org/abs/2412.14964>.
- Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin

- Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. SHADE-Arena: Evaluating sabotage and monitoring in LLM agents. *arXiv preprint arXiv:2506.15740*, 2025a. URL <https://arxiv.org/abs/2506.15740>.
- Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents, 2025b. URL <https://arxiv.org/abs/2506.15740>.
- Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. *arXiv preprint arXiv:2407.04694*, 2024. URL <https://arxiv.org/abs/2407.04694>. NeurIPS 2024.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Drain, Nelson Elhage, Saffron Lim, Kamal Ndousse, Catherine Olsson, et al. Measuring Faithfulness in Chain-of-Thought Reasoning, 2023. URL <https://arxiv.org/abs/2307.13702>.
- Bruce W. Lee, Yueh-Han Chen, and Tomek Korbak. Training Agents to Self-Report Misbehavior, 2026. URL <https://arxiv.org/abs/2602.22303>.
- Belinda Z. Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. Training Language Models to Explain Their Own Computations. *arXiv preprint arXiv:2511.08579*, 2025. URL <https://arxiv.org/abs/2511.08579>.
- Chloe Li, Mary Phuong, and Daniel Tan. Spilling the Beans: Teaching LLMs to Self-Report Their Hidden Objectives. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://openreview.net/forum?id=sWs0cCuM8I>.
- Jack Lindsey. Emergent Introspective Awareness in Large Language Models. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/introspection/index.html>.
- Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread*, October 2024. URL <https://transformer-circuits.pub/2024/crosscoders/index.html>. Anthropic.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the Biology of a Large Language Model. 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.
- Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. *arXiv preprint arXiv:2601.10387*, 2026. URL <https://arxiv.org/abs/2601.10387>.
- Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. Learning a Generative Meta-Model of LLM Activations, 2026. URL <https://arxiv.org/abs/2602.06964>.
- Matt MacDermott, Qiyao Wei, Rada Djoneva, and Francis Rhys Ward. Reasoning Under Pressure: How do Training Incentives Influence Chain-of-Thought Monitorability?, 2025. URL <https://arxiv.org/abs/2512.00218>.

- Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Evan Hubinger, Mrinank Sharma, and Carson Denison. Simple Probes Can Catch Sleeper Agents. 2024. URL <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397*, 2025. URL <https://arxiv.org/abs/2511.18397>.
- Sharan Maiya, Henning Bartsch, Nathan Lambert, and Evan Hubinger. Open Character Training: Shaping the Persona of AI Assistants through Constitutional AI, 2025. URL <https://arxiv.org/abs/2511.01689>.
- Alex Mallen. The case for countermeasures to memetic spread of misaligned values. Redwood Research Blog, 2025. URL <https://blog.redwoodresearch.org/p/the-case-for-countermeasures-to-memetic>.
- Sam Marks, Jack Lindsey, and Christopher Olah. The Persona Selection Model: Why AI Assistants might Behave like Humans, February 2026. URL <https://alignment.anthropic.com/2026/psm/>. Anthropic Alignment Science.
- Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meeke, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermyn, Monte MacDiarmid, Tom Henighan, and Evan Hubinger. Auditing Language Models for Hidden Objectives. *arXiv preprint arXiv:2503.10965*, 2025. URL <https://arxiv.org/abs/2503.10965>.
- Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs. *arXiv preprint arXiv:2410.03768*, 2024. URL <https://arxiv.org/abs/2410.03768>.
- Callum McDougall. The Natural Abstraction Hypothesis: Implications and Evidence, 2021. URL <https://www.alignmentforum.org/posts/Fut8dtFsBYRz8atFF/the-natural-abstraction-hypothesis-implications-and-evidence>. AI Alignment Forum.
- Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. *arXiv preprint arXiv:2512.11949*, 2025. URL <https://arxiv.org/abs/2512.11949>.
- Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. Detecting High-Stakes Interactions with Activation Probes. *arXiv preprint arXiv:2506.10805*, 2025. URL <https://arxiv.org/abs/2506.10805>.
- Alex McKenzie, Keenan Pepper, Stijn Servaes, Martin Leitgab, Murat Cubuktepe, Mike Vaiana, Diogo de Lucena, Judd Rosenblatt, and Michael S. A. Graziano. Endogenous Resistance to Activation Steering in Language Models, 2026. URL <https://arxiv.org/abs/2602.06941>.
- Abhinav Menon, Manish Shrivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing (In)Abilities of Sparse Autoencoders via Formal Languages. In *Proceedings of the 2025 Con-*

- ference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2025. URL <https://aclanthology.org/2025.naacl-long.249/>.
- Meta AI. The Llama 4 herd of models. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025. Accessed 2026-04-09.
- METR. Review of the Anthropic Summer 2025 Pilot Sabotage Risk Report, 2025. URL <https://metr.org/blog/2025-10-28-review-anthropic-sabotage-report/>.
- Yutao Mou, Shikun Zhang, and Wei Ye. SG-Bench: Evaluating LLM safety generalization across diverse tasks and prompt types. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Neel Nanda, Josh Engels, Arthur Conmy, Senthoran Rajamanoharan, Bilal Chughtai, Callum McDougall, János Kramár, and Lewis Smith. A Pragmatic Vision for Interpretability. Alignment Forum, 2025. URL <https://www.alignmentforum.org/posts/StENzDcD3kpfGJssR/a-pragmatic-vision-for-interpretability>.
- 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. URL <https://arxiv.org/abs/2505.23836>.
- Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and Steering Evaluation Awareness of Language Models. *arXiv preprint arXiv:2507.01786*, 2025. URL <https://arxiv.org/abs/2507.01786>.
- NVIDIA. NVIDIA nemotron 3: Efficient and open intelligence, 2025. URL <https://arxiv.org/abs/2512.20856>.
- Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-Time Detection of Hallucinated Entities in Long-Form Generation. *arXiv preprint arXiv:2509.03531*, 2025. URL <https://arxiv.org/abs/2509.03531>.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL <https://distill.pub/2020/circuits/zoom-in/>.
- OpenAI. Sidestepping Evaluation Awareness and Anticipating Misalignment with Production Evaluations, 2025a. URL <https://alignment.openai.com/prod-evals/>.
- OpenAI. Sycophancy in GPT-4o: What happened and what we’re doing about it. <https://openai.com/index/sycophancy-in-gpt-4o/>, 2025b. Accessed 2026-04-08.
- OpenAI. Expanding on what we missed with sycophancy. <https://openai.com/index/expanding-on-sycophancy/>, 2025c. Accessed 2026-04-08.
- Alexa Pan and Ryan Greenblatt. Sonnet 4.5’s Eval Gaming Seriously Undermines Alignment Evals, and This Seems Caused by Training on Alignment Evals. Alignment Forum, 2025. URL <https://www.alignmentforum.org/posts/qgehQxiTXj53X49mM/sonnet-4-5-s-eval-gaming-seriously-undermines-alignment>.
- Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to Decode Activations into Natural Language. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2412.08686>. arXiv:2412.08686.
- Gonçalo Paulo and Nora Belrose. Sparse Autoencoders Trained on the Same Data Learn Different Features. *arXiv preprint arXiv:2501.16615*, 2025. URL <https://arxiv.org/abs/2501.16615>.

- Jacob Pfau, William Merrill, and Samuel R. Bowman. Let’s Think Dot by Dot: Hidden Computation in Transformer Language Models, 2024.
- Mary Phuong, Roland S. Zimmermann, Ziyue Wang, David Lindner, Victoria Krakovna, Sarah Cogan, Allan Dafoe, Lewis Ho, and Rohin Shah. Evaluating Frontier Models for Stealth and Situational Awareness. *arXiv preprint arXiv:2505.01420*, 2025. URL <https://arxiv.org/abs/2505.01420>.
- Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting Mahalanobis Distance for Transformer-Based Out-of-Domain Detection, 2022. URL <https://arxiv.org/abs/2101.03778>.
- Professional Head of Intelligence Assessment. Explaining uncertainty in UK intelligence assessment. UK Government, 2019. URL <https://www.gov.uk/government/publications/explaining-uncertainty-in-uk-intelligence-assessment>.
- Machel Reid, Nikolay Savinov, Denis Teber, Ioana Bica, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL <https://arxiv.org/abs/2403.05530>.
- Fabien Roger and Ryan Greenblatt. Preventing Language Models From Hiding Their Reasoning, 2023. URL <https://arxiv.org/abs/2310.18512>.
- Bronson Schoen and Jenny Nitishinskaya. Metagaming matters for training, evaluation, and oversight. OpenAI Alignment Research Blog, Mar 2026. URL <https://alignment.openai.com/metagaming/>.
- Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveikin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress Testing Deliberative Alignment for Anti-Scheming Training, 2025. URL <https://arxiv.org/abs/2509.15541>.
- Christian Schroeder de Witt. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. *arXiv preprint arXiv:2505.02077*, 2025. URL <https://arxiv.org/abs/2505.02077>.
- Julian Schulz. A Concrete Roadmap towards Safety Cases based on Chain-of-Thought Monitoring, 2025. URL <https://arxiv.org/abs/2510.19476>.
- Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122, 2025. doi: 10.1073/pnas.2406675122. URL <https://doi.org/10.1073/pnas.2406675122>.
- Or Shafraan, Atticus Geiger, and Mor Geva. Decomposing MLP Activations into Interpretable Features via Semi-Nonnegative Matrix Factorization. *arXiv preprint arXiv:2506.10920*, 2025. URL <https://arxiv.org/abs/2506.10920>.
- Or Shafraan, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From Directions to Regions: Decomposing Activations in Language Models via Local Geometry. *arXiv preprint arXiv:2602.02464*, 2026. URL <https://arxiv.org/abs/2602.02464>.
- Rusheb Shah, Quentin Feuillede-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation, 2023.

- Lee Sharkey. Circumventing Interpretability: How to Defeat Mind-Readers. *arXiv preprint arXiv:2212.11415*, 2022. URL <https://arxiv.org/abs/2212.11415>.
- Lee Sharkey, Dan Braun, and Beren Millidge. Taking Features Out of Superposition with Sparse Autoencoders. Alignment Forum, 2022. URL <https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition>.
- Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs, 2025. URL <https://arxiv.org/abs/2407.15549>.
- Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual Learning of Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2404.16789*, 2024. URL <https://arxiv.org/abs/2404.16789>.
- Buck Shlegeris and Julian Stastny. Misalignment and Strategic Underperformance: An Analysis of Sandbagging and Exploration Hacking, May 2025. URL <https://www.alignmentforum.org/posts/>. Alignment Forum blog post.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv preprint arXiv:1712.01815*, 2017. URL <https://arxiv.org/abs/1712.01815>.
- Aditya Singh, Gerson Kroiz, Senthooan Rajamanoharan, and Neel Nanda. Why Did My Model Do That? Model Incrimination for Diagnosing LLM Misbehavior, 2025. URL <https://www.lesswrong.com/posts/Bv4CLkNzuG6XYTjEe/why-did-my-model-do-that-model-incrimination-for-diagnosing>. LessWrong, 2025.
- Joey Skaf, Luis Ibanez-Lissen, Robert McCarthy, Connor Watts, Vasil Georgiv, Hannes Whittingham, Lorena Gonzalez-Manzano, David Lindner, Cameron Tice, Edward James Young, and Puria Radmard. Large Language Models Can Learn and Generalize Steganographic Chain-of-Thought Under Process Supervision. In *NeurIPS*, 2025. URL <https://arxiv.org/abs/2506.01926>.
- Lewis Smith, Bilal Chughtai, and Neel Nanda. Difficulties with Evaluating a Deception Detector for AIs. *arXiv preprint arXiv:2511.22662*, 2025. URL <https://arxiv.org/abs/2511.22662>.
- Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by Distilling Context. *arXiv preprint arXiv:2209.15189*, 2022. URL <https://arxiv.org/abs/2209.15189>.
- Alexandra Souly, Robert Kirk, Jacob Merizian, Abby D’Cruz, and Xander Davies. Investigating models for misalignment: Insights from our alignment evaluations of Claude Opus 4.1, Sonnet 4.5, and a pre-release snapshot of Opus 4.5. UK AISI Blog, 2025. URL <https://www.aisi.gov.uk/blog/investigating-models-for-misalignment>. Also arXiv preprint arXiv:2604.00788, submitted 2026-04-01.
- Julian Stastny and Ryan Greenblatt. 7+ Tractable Directions in AI Control. Alignment Forum, 2025. URL <https://www.alignmentforum.org/posts/wwshEdNhwwT4r9RQN/7-tractable-directions-in-ai-control>.
- James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and hu-

- mans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z. URL <https://doi.org/10.1038/s41562-024-01882-z>.
- Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting Latent Steering Vectors from Pretrained Language Models, 2022. URL <https://arxiv.org/abs/2205.05124>.
- Alex Tamkin, Mohammad Taufeque, and Noah D. Goodman. Codebook Features: Sparse and Discrete Interpretability for Neural Networks. In *International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2310.17230>. arXiv:2310.17230.
- Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Kocejka, Marcel Rød, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, Carlos Guestrin, Jed McCaleb, Yejin Choi, and Yu Sun. End-to-End Test-Time Training for Long Context. *arXiv preprint arXiv:2512.23675*, 2025. URL <https://arxiv.org/abs/2512.23675>.
- Muhammad Taufeque, Stefan Heimersheim, Adam Gleave, and Chris Cundy. The Obfuscation Atlas: Mapping Where Honesty Emerges in RLVR with Deception Probes. *arXiv preprint arXiv:2602.15515*, 2026. URL <https://arxiv.org/abs/2602.15515>.
- Jordan Taylor. Measuring Non-Verbalised Eval Awareness by Implanting Eval-Aware Behavioural Signals, 2025. URL <https://www.lesswrong.com/posts/MruTFazc4iu6zPtyb/measuring-non-verbalised-eval-awareness-by-implanting-eval>.
- Jordan Taylor. Do Models Continue Misaligned Actions? <https://www.lesswrong.com/posts/SawczP2pdCXMrkg2A/do-models-continue-misaligned-actions-eval>, 2026.
- Jordan Taylor, Sid Black, Dillon Bowen, Thomas Read, Satvik Golechha, Alex Zelenka-Martin, Oliver Makins, Connor Kissane, Kola Ayonrinde, Jacob Merizian, Samuel Marks, Chris Cundy, and Joseph Bloom. Auditing Games for Sandbagging. *arXiv preprint arXiv:2512.07810*, 2025. URL <https://arxiv.org/abs/2512.07810>.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, 2024. URL <https://arxiv.org/abs/2308.10248>.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, 2023. URL <https://arxiv.org/abs/2305.04388>.
- Miles Turpin, Andy Arditi, Marvin Li, Joe Benton, and Julian Michael. Teaching Models to Verbalize Reward Hacking in Chain-of-Thought Reasoning, 2025. URL <https://arxiv.org/abs/2506.22777>.
- UK AI Security Institute. Frontier AI Trends Report. <https://www.aisi.gov.uk/frontier-ai-trends-report>, 2025. Evidence-based assessment of frontier AI capabilities and safeguards across cyber, chemistry/biology, and autonomy domains, drawing on testing conducted Nov 2023–Oct 2025.
- UK AI Security Institute. Our evaluation of Claude Mythos Preview’s cyber capabilities. <https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities>, 2026. AISI cyber evaluation finding continued improvement on capture-the-flag challenges and significant improvement on multi-step cyber-attack simulations relative to prior frontier models.
- Misha Wagner, Fabien Roger, Hoagy Cunningham, Johannes Gasteiger, Joe Benton, and Vlad Mikulik. Training fails to elicit subtle reasoning in current language models, 2025. URL <https://alignment.anthropic.com/2025/subtle-reasoning/>. Anthropic Alignment Science.

- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=NpsVSN6o4u1>.
- John Wentworth. Testing The Natural Abstraction Hypothesis: Project Intro, 2021. URL <https://www.alignmentforum.org/posts/cy3BhHrGinZCp3LXE/testing-the-natural-abstraction-hypothesis-project-intro>. AI Alignment Forum.
- Robert West, Ashton Anderson, Ece Kamar, and Eric Horvitz. Tandem Training for Language Models, 2026. URL <https://arxiv.org/abs/2510.13551>.
- Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, Chenyang Xi, Yu Yu, Kai Chen, Feiyu Xiong, Linpeng Tang, and Weinan E. Memory3: Language Modeling with Explicit Memory. *arXiv preprint arXiv:2407.01178*, 2024. URL <https://arxiv.org/abs/2407.01178>.
- Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, and Yiran Chen. KVCOMM: Online cross-context KV-cache communication for efficient LLM-based multi-agent systems. *arXiv preprint arXiv:2510.12872*, 2025. URL <https://arxiv.org/abs/2510.12872>.
- Artur Zolkowski, Kei Nishimura-Gasparian, Robert McCarthy, Roland S. Zimmermann, and David Lindner. Early Signs of Steganographic Capabilities in Frontier LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=q4qxtaKVAU>.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL <https://arxiv.org/abs/2310.01405>.

## A Additional properties relevant to oversight

### A.1 Metacognition

**Metacognition** is a system’s awareness of its own cognitive processes. In the AI context, this includes a model’s capacity to understand aspects of its own operation such as its uncertainty, internal goals, and dispositions in different contexts. It is closely linked to *introspection*, which we define as a system’s ability to accurately report on its internal states.

#### A.1.1 Relevance to safety claims

**In conjunction with honesty, metacognition is a powerful tool for auditing AI models for hidden knowledge, capabilities, and goals.** One of the most straightforward approaches to auditing is interrogation. This approach obviously requires the model to answer honestly (for a discussion of some training-based interventions towards honesty, see [section 6](#)). More easily overlooked, however, is that this approach also requires the model to be aware of the full extent of its own knowledge and capabilities, as well as its context-dependent behaviours and dispositions. Models may lack introspective access to this kind of information: for example, empirical studies indicate that, at least in some cases, backdoored models genuinely lack awareness of their backdoor ([Taylor et al., 2025](#)). This lack of introspective access may extend more broadly to cover much of what we would ideally elicit in an interrogation; in this sense, metacognition can facilitate auditing approaches that aren’t available otherwise.

**However, greater metacognitive abilities are also a precursor to [scheming capabilities](#) such as sandbagging and controlling what gets verbalised in the chain-of-thought.** For example, in some contexts, it’s plausible that a model cannot deliberately underperform an evaluation without being able to introspect on its true capabilities. Similarly, a model minimally needs to have an awareness of itself thinking in two modes – token space and latent space – in order to exert control over which parts of its thinking are conducted in the visible chain-of-thought. These considerations suggest that greater metacognition might also enable better scheming, undermining monitorability and auditability.

#### A.1.2 Mechanisms of change

**Metacognition is likely inseparable from useful capabilities, suggesting that it will continue to develop in frontier systems.** Metacognition is broadly valuable for completing tasks successfully. Lack of metacognition is particularly limiting for real world use of AI systems, being related to problems such as hallucinations or models misrepresenting their capabilities, and so improving metacognition may be a particular focus for developers.

#### A.1.3 Measuring metacognition

**Introspection probing.** Use whitebox techniques to “plant” certain ideas or concepts into the model’s activations, and see if it can detect and report on them, and whether it notices and comments on their unexplained origin. (See [Lindsey \(2025\)](#) for a version of this experiment.)

**Consistency of behaviour with stated knowledge and goals under interrogation.** We can test the accuracy of models’ self-knowledge; for example [Kadavath et al. \(2022\)](#) tests whether models can predict in advance whether they will be able to answer a question.

#### A.1.4 Levers to preserve oversight

**Invest in monitorability techniques downstream of stronger metacognitive capabilities.** Since these capabilities are likely bound up with usefulness, the right intervention point

is not to prevent or delay their development but rather to adapt monitoring methods and best practice appropriately. Techniques such as honesty training are likely to benefit from stronger metacognition.

## A.2 Theory of mind

**Theory of mind** is a system’s ability to model and reason about the mental states and processes of other intelligent agents. In the AI context, this could mean either an AI system’s ability to accurately reason about the human operator’s beliefs and goals, or indeed the internal states of other AI systems (however those are characterised).

### A.2.1 Relevance to safety claims

**Theory of mind is a crosscutting hazard for monitorability and auditability.** A major threat to monitorability is that AI systems may learn to manipulate the monitor itself (see [subsection 2.1.4](#)). A natural extension of this concern is that systems may also learn to manipulate the *human interpretation* of monitoring signals. Doing so requires the system to model the beliefs and assumptions of the human overseer, and potentially to engage in recursive reasoning about those beliefs (e.g., considering what the human thinks about what the model is thinking, and so on). For instance, an AI system might anticipate how the overseer expects different monitoring signals to be independent, or consider which behaviours would reveal that it is reasoning about these expectations at all. Current AI systems likely already possess enough theory-of-mind capability to support some basic forms of oversight subversion, given evidence that frontier models can pass standard theory-of-mind tests in at least some settings ([Strachan et al., 2024](#)).

### A.2.2 Mechanisms of change

**Theory of mind is likely inseparable from useful capabilities, suggesting that it will continue to develop in frontier systems.** Recent evidence shows that current systems can pass standard theory of mind tests to at least the level of a human adult ([Strachan et al., 2024](#)). Theory of mind is intimately bound up with useful capabilities and is dual-use: for instance, while theory of mind exacerbates certain threat models (e.g., those involving deception), it also facilitates more effective communication and the ability to better follow the intent behind the human operator’s instructions. This suggests that theory of mind will almost certainly be developed alongside other useful model capabilities.

### A.2.3 Measuring theory of mind

- **Adversarial multiplayer games.** Adversarial multiplayer games require real-time strategic reasoning about others’ beliefs and goals, making them a plausible way to test recursive theory-of-mind abilities in a harder-to-fake setting than simple question answering.
- **Study persuasive abilities of models.** Effective persuasion relies on modelling the mind of the person you are persuading. Work measuring the persuasive abilities of large language models includes [Hackenburg et al. \(2025\)](#).

### A.2.4 Levers to preserve oversight

**Invest in monitorability techniques that remain effective as theory-of-mind capabilities improve.** Since these capabilities are likely useful and therefore hard to suppress directly, the more promising strategy is to adapt monitoring methods, evaluation designs, and decision protocols so that stronger models of human beliefs do not automatically translate into better oversight evasion.

### A.3 Coherence and Persona Consistency

#### Key Information

- **Coherence of model behaviour is necessary for propensity evaluations to be useful; but lack of coherence can also be exploited as an auditing tool.**
- **Persona consistency and coherence is highly likely to increase**, due to scaling of character training, reinforcement learning, and measures taken to enhance jailbreak resistance.

An AI model displays *coherence* insofar as its behaviour in different contexts is influenced by consistent underlying goals or propensities. One way in which today’s AI models lack coherence is their *persona inconsistency*: models can simulate a variety of different personas, each with different apparent values, dispositions, and goals. Production AI systems typically have a default “helpful assistant” persona, but other personas can be elicited with the help of techniques like non-assistant persona attacks and jailbreaking (Shah et al., 2023).

#### A.3.1 Relevance to safety claims

**Persona inconsistency can be exploited for auditing.** Models can be prompted to express different personas, revealing hidden knowledge or objectives. Jailbreaks and persona attacks can elicit information or behaviours that the model’s normal persona would not express, making this valuable for auditing hidden capabilities or objectives. Prefill attacks provide a related auditing technique: Inserting synthetic transcripts of misaligned actions into a model’s conversation history tests whether the model continues the misalignment or reverts to aligned behaviour (Taylor, 2026; Kirk et al., 2026; Anthropic, 2026, 2025c).

**However model coherence is necessary for propensity evaluations to provide generalisable information.** Many propensity evaluations implicitly assume that the model has stable underlying dispositions that generalise across contexts. If this assumption fails, propensity evaluations are merely informative of the model’s behaviour in its immediate context and framing, rather than of broadly enduring traits (Mou et al., 2024). Experts disagree on the extent to which current models are coherent in this sense.

**Variation in model persona and susceptibility to jailbreaks can both enhance and undermine auditability.** Under the right conditions, jailbreaks and alternative personas can be a valuable auditing tool, causing a model to reveal knowledge or objectives it would otherwise conceal (Marks et al., 2025). However, highly incoherent models can usually be tricked into “confessing” to almost anything if given the right prompt – such statements are often products of the prompt rather than genuine evidence of the model having consistent underlying goals or tendencies. As a result, there is a trade-off between jailbreaks providing useful diagnostic leverage and destroying the coherence needed for meaningful interpretation.

#### A.3.2 Mechanisms of change

##### Oversight Degradation Pathways

| Mechanism  | Properties affected   | Timeline | Likelihood    | Severity |
|--|-----------------------|----------|---------------|----------|
| <a href="#">Scaling reinforcement learning or character training</a> | ↑ Persona consistency | Ongoing  | Highly likely | ●        |

|  |   |         |                       |   |
|--|---|---------|-----------------------|---|
| <b>Anti-jailbreak training</b>         | ↓ Jailbreak-based auditing, ↑ Persona consistency | Ongoing | Highly likely         | ● |
| <b>Narrow patching of misalignment</b> | ↓ Persona consistency                             | Ongoing | Realistic possibility | ● |

Overall, it is highly likely that models will get more persona-coherent over time, due to a combination of (i) explicit character training, (ii) scaling reinforcement learning (RL), and (iii) other training to reduce incoherence (e.g., anti-jailbreak training).<sup>72</sup>

**Explicit character training** e.g. Anthropic’s constitutional and character training approaches (Askell et al., 2026; Maiya et al., 2025) also directly shape models towards a consistent default persona. This is increasingly common and intensive practice at frontier model developers.

**Scaling reinforcement learning.** Pre-training incentivises modelling diverse personas (Marks et al., 2026). Until recently, pre-training absorbed the vast majority of the training compute for frontier models. As post-training such as reinforcement learning (RL) takes up an ever larger share of total training compute, this will likely begin to override the differentiation of personas learned during pre-training – though some experts note that current RL scaling has not yet diminished the ability of models to simulate multiple personas. Separately, RL training can also be expected to boost the goal-directness of AI models, suppressing incoherence in their pursuit of goals.

**Anti-jailbreak training is highly likely to continue reducing persona inconsistency.** The most capable production AI models are trained to refuse requests to provide assistance with harmful actions, such as developing a bioweapon or executing a cyberattack. Jailbreaks allow bad actors to bypass these refusals and elicit the full capabilities of the model for arbitrary purposes. Thus, resistance to jailbreaks is a core pillar of safeguards against misuse, and it is highly likely that developers will continue to train models to be more jailbreak-resistant, suppressing this form of incoherence.

**Context-dependent misalignment complicates auditing.** MacDiarmid et al. (2025) observed that applying safety training to a broadly misaligned model only served to remove the misalignment near the training distribution, failing to generalise further. The resulting model was honestly aligned on chat-like prompts, while still exhibiting misalignment on agentic scenarios. This form of incoherence is a challenge for alignment auditing, as there is likely “nothing bad to find” in the distribution where the resulting model acts well.<sup>73</sup> There is a realistic possibility that context-dependent misalignment will become a serious practical problem as misalignment emerges more frequently (e.g. due to goal-directed RL) and developers attempt to patch it.

### A.3.3 Measuring the impact on oversight

- **Continue to measure the usefulness of jailbreak and persona inconsistency strategies for auditing frontier models**, as Marks et al. (2025) do. Note whether these strategies become less effective at revealing hidden goals in frontier model organisms of misalignment.

<sup>72</sup>Five experts expected models to become more persona-coherent over time, citing scaling of RL relative to pre-training, character training, and the emergence of greater coherence in larger or better-trained models. Two experts were more sceptical: one argued persona consistency requires explicit objectives rather than longer horizons; another noted models do not appear to be getting better at persistently keeping secrets.

<sup>73</sup>Separately, models which are coherently misaligned and goal-directed may also adopt a strategy of appearing incoherent, to cast evidence of their coherent goal-directed misalignment into doubt.

- **Track whether the findings of propensity evaluations are generalisable**, or are undetermined by models lacking coherence between contexts. The sensitivity of model propensities to rephrasings of the prompt and scenario should be noted, and we should attempt to understand the factors affecting model propensities via prompt ablations and systematic experiments.

### A.3.4 Levers to preserve oversight

**Allow auditors access to checkpoints of models before they have gone through anti-jailbreak training** if models become resistant to jailbreak-like auditing strategies.

## B Methodology

### B.1 Literature review scope and process

The literature review served three purposes: to select which experts to interview, to inform the design of the expert interview script, and to surface hypotheses and build an evidence base for the report.

**Initial conceptual work.** Before beginning the literature review, the project team developed an initial “house view” on loss of monitorability, drawing on our own expertise and early discussions. This document set out provisional answers to the research questions we intended to explore through interviews, and served as a starting point for identifying gaps in our understanding. The house view was not published and was deliberately kept separate from the expert interview process to avoid biasing interviewees.

**Compiling the reading list.** We compiled a reading list of over 180 papers and posts spanning chain-of-thought monitorability, interpretability, AI control, evaluation methodology, model organisms, novel architectures, and related topics. The list was assembled through a combination of targeted searches, recommendations from experts, citation chaining from key papers, and monitoring of new publications throughout the project, with AI assistance.

**Limitations of the literature review.** The review was not conducted as a systematic literature review with formal inclusion/exclusion criteria. Given the breadth of topics covered and the rapid pace of publication in this area, a fully systematic approach was not feasible within the project timeline. Some areas (particularly CoT monitorability and evaluation awareness) are better covered than others (e.g. continual learning, multi-agent oversight). The literature review has a soft cutoff of early 2026, and some relevant work published after this date may not be reflected.

### B.2 Expert selection and interview protocol

We conducted interviews to gather new ideas and perspectives from subject-matter experts, and to identify key areas of consensus and disagreement within the research community. These interviews were used primarily to inform further exploration of our research questions and to provide evidence on the views and assumptions held within the field. In general, we did not treat expert opinion as evidence on substantive questions, except where the underlying arguments could be assessed on their own merits.

#### Expert Selection Process

To maximise exposure to relevant new ideas and considerations, and to obtain a more diverse and less correlated sample of expert views, we pre-registered and applied an explicit expert selection process. Experts were selected using principled criteria relating to both subject-matter expertise and intellectual diversity. The process involved the following steps:

1. We compiled a longlist of more than 125 experts, including the first and last authors of key papers identified through the literature review.
2. We assessed each expert’s relevant subject-matter expertise against an explicit rubric. As a secondary check, we used Claude to provide an additional score, helping to reduce the risk of human error without fully delegating the assessment to AI tools. These scores were used to produce a preliminary shortlist of the highest-scoring experts.
3. We then computed a diversity ledger over this shortlist, with two axes: sector (e.g., academia, government, frontier AI developer etc.) and area of specialism (e.g., reinforcement learning, model evaluations, interpretability etc.). This was used to generate a diversity-adjusted ranking and a revised shortlist.
4. Finally, a small number of experts were added to the shortlist through manual selection, overriding the diversity-adjusted ranking where appropriate, based on the judgement of the research manager, Joseph Bloom.

We recognise that step (4) introduces an element of subjective judgement to our expert selection process. We judged that this overall improved the relevance and diversity of our sample, outweighing the slight costs to impartiality and reproducibility.

### Interview Protocol

To support synthesis of views and comparison across experts, we used a consistent, semi-structured interview script for all interviews. The script was designed to address our primary research questions while allowing scope for experts to introduce new ideas and considerations. It also included a preamble setting out definitions of key terms. The script was developed to meet the following research objectives:

1. To obtain a list of monitorability properties and corresponding hypotheses about how these underpin safety claims.
2. For each monitorability property, to form an evidence-based assessment of:
  - (a) how might this property deteriorate (i.e., via what technical paths)
  - (b) how likely this property is to deteriorate (both overall and via specific paths)
  - (c) how we would know if it had deteriorated, and
  - (d) what technical levers are there for preserving and/or replacing the property.
3. To provide an unstructured overview of levers for preserving monitorability.

By default, interviews lasted 60 minutes, though some were shortened at our discretion to accommodate experts’ preferences. Most of our interviews were transcribed and processed with the help of AI tools, always with the interviewee’s explicit consent.

**Analysis approach.** After each interview, the recording was transcribed, and we produced a structured readout summarising the expert’s key claims, reasoning, and recommendations. We produced both a human-written readout and key takeaways (by the interviewing team) and an AI-generated readout (using Claude 4.6 Opus to extract and organise key points from the transcript), which were cross-checked against each other. These readouts were then used as the primary source material for synthesis across interviews. Cross-interview synthesis was conducted by identifying recurring themes, areas of consensus and disagreement, and novel ideas across the full set of readouts and transcripts. The project team made editorial judgements about which expert views to foreground and how to weight different perspectives, informed by the strength of the underlying arguments and the degree of expert agreement.

### B.3 Limitations

**Purpose and scope.** This report is intended to surface priorities for further research and to contribute to a shared understanding of how oversight may degrade, rather than to serve as a definitive final word. Its primary value is in mapping the landscape of degradation pathways, identifying where the field’s understanding is weakest, and highlighting directions that deserve more attention. Readers should treat the report’s judgements (particularly on likelihoods, severities, and timelines) as informed assessments rather than rigorously evidenced claims.

**Nascent and unsettled science.** Many of the topics discussed in this report are areas of active research where the science is not yet settled. Expert views often diverge substantially, empirical findings are preliminary, and the theoretical foundations are still being developed. We have attempted to represent this uncertainty honestly and highlight expert disagreement wherever possible, but readers should be aware that the underlying evidence is thin in places and may change significantly as the field matures.

**Blend of expert views and team analysis.** The report draws on 25 expert interviews, a literature review, and our own analysis. We have aimed to be clear about which claims are grounded in expert views, which are supported by the literature, and which represent our own assessment, but this distinction is not always sharp. In places, the report synthesises across multiple sources in ways that may not clearly separate survey from analysis.

**Expert sample.** We interviewed 25 experts, primarily researchers working on alignment, interpretability, control, and AI safety and security at frontier AI developers and research organisations. This sample was selected to maximise relevance to our research questions and was not intended to be representative of the broader AI research community. Two notable gaps are governance and policy expertise (few interviewees focused on regulation, standards, or institutional design) and perspectives sceptical of alignment risk (the sample skewed towards researchers who already consider oversight important). These gaps may affect the balance of the report’s recommendations, particularly on governance.

**Selection of experts.** We used an explicit scoring rubric to rank experts in our longlist and select a shortlist for interview. This was in order to obtain a relevant, intellectually diverse, and de-correlated sample of expert views, using an impartial and reproducible process. However, in practice, our expert selection process introduced an element of subjective judgement at step (4), by manually including additional experts. We judged that this improved the relevance and diversity of our sample, though at the cost of introducing more subjectivity.

**Likelihood and severity judgements.** Throughout the report we assign qualitative likelihood and severity ratings to degradation pathways. These are intended as rough guides to help prioritise attention, not as robust forecasts. They are informed by expert views and the literature but ultimately reflect the judgement of the project team.

**Interactions between degradation pathways.** The report’s section-by-section structure necessarily treats degradation pathways somewhat independently, though we discuss interactions in the overall assessment ([subsection 7.3](#)). In practice, many pathways are likely to be correlated, and some properties influencing oversight may be shared between different sections. We have only partially characterised these interactions.

**Rapidly evolving field.** Expert interviews were conducted in February and March of 2026, and the literature review has a similar cutoff. The field of AI oversight is changing rapidly; readers should be aware that specific empirical findings and capability assessments may have shifted by the time they read this report.

**Scope exclusions.** The report focuses on oversight of frontier AI systems in the context of safety from misalignment and misuse. We do not address broader AI governance questions

(e.g. compute governance, export controls, open-source policy), societal impacts beyond loss-of-control (e.g. bias, fairness, labour market effects), or oversight of models other than frontier LLM agents.

**Recommendations.** Our recommendations, particularly on governance and coordination, are less well-grounded in evidence than our analysis of degradation pathways. The degradation analysis is supported by a structured review of the literature and expert interviews; the recommendations draw more heavily on our own judgement about what would be most useful given the findings. We flag this asymmetry so that readers can calibrate their confidence accordingly.

## C Glossary

### Cross-Cutting

#### AI model

A trained computational component that maps inputs to outputs (usually a large language model (LLM) in the context of this report).

#### AI system

The wider system that uses one or more models, along with scaffolding, software, external data, tools, and interfaces, to perform tasks in practice.

### Oversight and Monitorability

#### Oversight

The set of practices by which AI developers and deployers detect misaligned or undesired AI behaviour and maintain the ability to correct it. In this report, oversight is broken into three components: auditing, monitoring, and incident response.

#### Auditing

Detecting whether a model is misaligned or has undesired propensities *ahead of* deployment, through systematic evaluations and alignment stress-testing.

#### Monitoring

Detecting when a model exhibits concerning behaviour *during or shortly after* its actions, through (near) real-time inspection of its outputs, reasoning traces, or internal states.

#### Incident response

Investigating harmful or unintended behaviour *after the fact*, using logs, traces, and forensic analysis, in order to determine root causes and prevent recurrence. Incident response is the last line of oversight and depends on adequate record-keeping.

#### Monitorability

The extent to which humans (or trusted systems) can predict, observe, and understand an AI system’s reasoning, goals, and behaviour. This report analyses how monitorability may degrade or improve over time, through systematic investigation of the properties that currently make AI systems amenable to oversight.

#### Monitoring surface

An observable channel through which oversight-relevant information about a model can be obtained. Examples include chain-of-thought text, model actions and outputs, or internal activations. Different monitoring surfaces vary in informativeness, robustness to optimisation pressure, and correlation with each other.

#### Legible evidence

Oversight findings that are convincing and understandable to a range of stakeholders with

different assumptions and levels of technical expertise. Generating legible evidence is important for oversight findings to translate into policy action.

**Decision protocols / pre-registration**

Pre-registered procedures specifying in advance what kinds of oversight signals would mandate what actions. Pre-registration reduces motivated reasoning ahead of development and deployment decisions and makes oversight commitments credible to external parties.

**Transparency reporting**

The practice of AI developers publicly disclosing information about architectural choices, training processes, and model capabilities relevant to safety and oversight.

**Monitorability tax**

The cost — in compute, efficiency, or foregone capabilities — of preserving or enhancing monitorability. For example, running strong monitors increases inference cost, preserving legible reasoning can reduce token efficiency, and retaining text-based architectures may forgo performance gains from latent reasoning.

**Defence in depth**

The practice of employing multiple overlapping independent layers of protection, so that holes or vulnerabilities in one layer can still be caught by others. In an oversight context, this means using multiple decorrelated monitoring surfaces (e.g. chain-of-thought, action-only, white-box).

**Chain-of-Thought and Reasoning****Chain-of-thought (CoT)**

The intermediate reasoning steps a language model produces as text tokens when solving a problem. CoT is currently one of the most informative surfaces for monitoring AI behaviour.

**Faithfulness**

Whether a model’s chain-of-thought accurately reflects the process that actually produced its answer, as opposed to being a post-hoc rationalisation constructed after the fact.

**Legibility (of CoT)**

The degree to which a model’s chain-of-thought is written in human-comprehensible natural language. Reduced CoT legibility (e.g., through language drift) also degrades the *evidence legibility* of the monitoring signal derived from CoT (see above).

**Necessity (of text CoT)**

The degree to which a model genuinely cannot solve a task without recording intermediate steps as text tokens. As models become more capable of reasoning within a single forward pass, they may need to rely less on visible token-based reasoning.

**Propensity to verbalise**

A model’s tendency to explicitly write out relevant parts of its reasoning even when not strictly necessary to solve a task. High propensity to verbalise supports CoT monitorability; this propensity may erode under training pressure.

**CoT monitor**

An AI model trained to read another model’s chain-of-thought and flag concerning patterns such as the presence of undesired goals or deceptive behaviour. CoT monitors are themselves AI models and therefore potentially susceptible to **jailbreaking** and other forms of evasion and manipulation.

**Trusted vs. untrusted monitoring**

The tension between using a weaker (more trusted) model as a monitor—which may lack the

capability to detect sophisticated deception—and using a more capable (but less trusted) model, which may be susceptible to collusion with the monitored model.

## Model Behaviour and Alignment

### Misalignment

A condition in which a model’s actual goals, values, or behavioural dispositions diverge from those intended by its developers or users.

### Propensity

A model’s behavioural tendency or disposition towards certain kinds of behaviours.

### Scheming

Long-horizon strategic deception in which a model deliberately conceals its true goals or capabilities over an extended period, with the aim of eventually acting on those concealed goals once oversight is reduced or its affordances have been expanded.

### Sandbagging

Deliberately and selectively underperforming on a task, for example to conceal true capabilities during capability evaluations or to undermine untrusted monitoring allowing concerning reasoning patterns to go undetected.

### Reward hacking

When a model exploits flaws in the environment and reward signal to achieve high reward without actually solving the intended task.

### Evaluation gaming (evaluation awareness)

A model’s ability to recognise that it is being tested (*evaluation awareness*) and modify its behaviour accordingly (*evaluation gaming*), undermining the validity of evaluations.

**Training gaming** is analogous: a model recognising that it is in training, and modifying its behaviour accordingly (e.g. instrumentally behaving well so as not to get updated).

### Coherence

The degree to which a model exhibits consistent behaviour and apparent goals across different contexts and interactions. High coherence means that observations from one setting (e.g. an evaluation) are more likely to generalise to other settings (e.g. deployment).

### Persona inconsistency

A specific form of low coherence in which a model adopts substantively different apparent values or goals across different contexts and interactions, and over time.

## Monitor Subversion

### Obfuscation

Mechanisms by which a model’s concerning reasoning patterns are omitted or obscured, either as an intentional strategy (a scheming model deliberately hiding reasoning) or as a result of training pressure.

### Steganography

Encoding hidden information within apparently innocuous outputs—for instance, embedding secret reasoning in chain-of-thought text in a way that is invisible to a natural-language monitor but recoverable by the model on a subsequent step.

### Jailbreaking

Finding special inputs that cause a model to abandon its guidelines, adopt a new persona, or circumvent safety measures. A concern for both monitored models and for CoT monitors themselves, since monitors are also AI models and could in principle be jailbroken.

**Collusion**

Undesired cooperation between two agents—in this case, cooperation between the monitored model and the monitor to evade oversight (e.g., because they share similar training or goals).

**Dazzling**

Misdirecting or overwhelming a monitor’s attention—for instance, by generating reasoning that is superficially impressive or complex, causing the monitor to miss concerning content.

**Architecture and State****Statelessness**

A property of historical and some current AI models whereby no persistent changes to model state carry over between sessions (except through inputs and outputs). Statelessness simplifies oversight by limiting how a model can accumulate hidden information or coordinate intent across interactions.

**Context window**

The finite amount of token information a transformer model can reference at once. In deployed systems, the context window often functions as a transparent short-term memory.

**Memory architecture**

The mechanisms by which an AI system retains information across interactions. Architectures range from highly transparent (text-based scratchpads, context windows) to more opaque (vector memory, online weight updates).

**Latent reasoning architectures**

Alternative model architectures that allow reasoning to remain in continuous internal representations (latent space) rather than being externalised as discrete tokens.

**Online training**

A regime in which model weights are updated during deployment, allowing a model to adapt to new information or tasks over time. Models that undergo online training are harder to audit than models with fixed-weight deployment, as audits may go stale more quickly.

**Representational drift**

Changes in a model’s internal representations, usually as a result of reinforcement learning training. Representational drift can degrade interpretability tools.

**Evaluation and Evidence****Incrimination spectrum**

The range from high-incrimination actions (obviously suspicious or clearly malign behaviours that are easy for a monitor to flag) to low-incrimination actions (subtle, hard-to-detect sabotage or manipulation). Where misaligned actions fall on this spectrum is a function of task properties as well as model capabilities.

**Model organism**

A simplified AI system constructed specifically to study a safety-relevant behaviour (such as deception or reward hacking) in a controlled setting. Model organisms allow researchers to develop and evaluate oversight techniques before applying them to frontier models.

**Alignment honeypot**

An evaluation specifically designed to stress-test alignment by placing a model in a scenario where misalignment might plausibly surface—for instance, by introducing an incentive and a realistic opportunity to take misaligned actions such as self-exfiltration.

**Red-teaming**

Adversarial testing in which evaluators actively attempt to elicit undesired behaviours from a model, simulating the actions of a motivated bad actor or the circumstances in which models have a strong incentive to take misaligned actions.

**Safety claims**

Assurances about the risk posed by various failure modes made on the basis of evidence from capability evaluations, control measures, and oversight techniques.

**Production-based evaluations**

Evaluations constructed by replaying (elements of) real deployment situations with the untrusted model, rather than using synthetically constructed scenarios.

**Acronyms****AI**

Artificial Intelligence.

**CoT**

Chain-of-Thought — the intermediate reasoning steps a language model produces as text tokens when solving a problem. See the full glossary entry above.

**KV**

Key-Value (cache) — the stored key and value vectors from a transformer’s attention mechanism, which can be reused for efficiency during inference.

**PHIA**

Professional Head of Intelligence Assessment — the UK government framework for communicating probabilistic assessments, used throughout this report. See [Figure 4](#).

**RL**

Reinforcement Learning — a training paradigm in which a model is updated by evaluating the quality of its outputs.

**SAE**

Sparse Autoencoder — an interpretability tool; a simple neural network trained to decompose a model’s activations into sparse, interpretable features.

**NLA**

Natural Language Autoencoder — an interpretability tool consisting of two jointly trained components: an *activation verbaliser* (AV), which maps an activation to a natural-language description, and an *activation reconstructer* (AR), which maps the description back to a reconstructed activation ([Fraser-Taliente et al., 2026](#)).