

# 18-month update Apollo Research

**Apollo Research** is an evaluation organization focusing on risks from deceptively aligned AI systems. We conduct technical research on AI model evaluations and interpretability and have a small AI governance team. As of December 2024, we are 18 months old.

## Executive Summary

We worked with OpenAI to test o1 before public deployment. Alongside [OpenAI's system card](#), we released [a paper](#) with detailed explanations and additional results. Our findings have been widely covered in the media, e.g. [TIME](#), [The Times](#), [TechCrunch](#), [The Information](#), and more.

The **evaluations** team is working on [scheming-related capability evaluations](#), [safety cases](#), and [pre-deployment testing of frontier models](#). In 2025, we plan to increase these efforts while also focusing on automating the evals pipeline. Our goal is to help governments and AI developers understand, assess, and address the risks of deceptively aligned AI systems.

The **interpretability** team is working on a new mechanistic interpretability method that potentially overcomes the issues of previous methods. We have some encouraging early results and intend to publish a draft write-up in early 2025. Additionally, we're applying existing white-box approaches to detect deception in LLMs and will publish our findings in early 2025. We're also working with a large philanthropist to increase the funding available to the field.

The **governance** team's goal is to support important governance developments, such as the EU's Code of Practice with tailored research and advice. Next, we will continue efforts to build an evaluation ecosystem and start research on the governance of internal deployment.

In the last six months, we have expanded several existing partnerships and launched new noteworthy partnerships. We were invited to the AISI Network Meeting in SF (Nov. 2024) and the French AI Action Summit (Feb. 2025). We expanded our partnerships with UK AISI and frontier AI labs like OpenAI, Anthropic, and Google DeepMind and are beginning new partnerships with Amazon and Microsoft.

Over the last six months, we hired five more full-time staff. In all cases, job offers were accepted by our top choices of candidates and the seniority of our applicants increased, which we see as positive signs.

While we have resolved most of the challenges we laid out in our 1-year report, we are still looking for recurring 7-figure long-term commitments from philanthropic funders, and a lot of leadership time is spent on fundraising.

We have more room for funding. Please [reach out](#) if you're interested.

# Selection of completed and future work

## Evaluations

We think about **scheming safety cases** as the conceptual foundation of our evals work. In our paper on [evaluations-based safety cases for scheming](#), we suggested three core arguments—scheming reasoning inability, scheming harm inability, and scheming control. We plan to continue thinking about safety cases for scheming, both refining existing arguments as well as designing new ones, e.g. by investigating the limitations of latent reasoning of LLMs.

Until ~November 2024, we have primarily focused on building **evaluations for scheming reasoning inability** arguments. This has resulted in our paper “[Frontier Models are Capable of In-context Scheming](#)”. We found that multiple frontier models are capable of in-context scheming and, therefore, think that a general scheming inability argument is not applicable anymore. Thus, we recommend that the chain-of-thought of the most capable models should be monitored for evidence of scheming.

Over the next six months, we will shift our priorities to designing **evaluations for scheming harm**. We specifically focus on the ability of AI systems to sabotage AI research or other AI systems in an internal deployment setting. For example, we want to evaluate whether current models are capable of training a complex backdoor into another AI system or strategically influencing the research of an AI safety researcher.

Starting in ~Q2 2025, we intend to shift focus to **scheming control evaluations**, i.e. designing and evaluating control protocols for frontier AI systems that reduce risk under the worst-case assumption that the model is competently scheming. These protocols will be especially important when competent AI agents are able to take over meaningful research tasks.

We are **working with frontier AI companies** in various ways, including running pre-deployment evaluations to inform their deployment strategy, developing evaluations with them, and advising on their safety frameworks.

Beyond designing and running evaluations, we have found early success with using LLMs for better evals tooling to substantially speed up the development and execution workflow. We want to continue to think about how we can **automate larger parts of the evals pipeline**.

## Interpretability

Mechanistic interpretability traditionally follows a two-step approach: decomposing neural networks into components, then characterizing their functions. While Sparse Dictionary Learning has emerged as the primary decomposition method for identifying features in neural activations, recent work has highlighted both practical and theoretical limitations (e.g. [Bussman et al. 2024](#);

[Engels et al. 2024a](#), [Engels et al. 2024b](#)). Importantly, the ability to identify features does not immediately tell us how the network's parameters compute them.

We are, therefore, working on a new mechanistic interpretability approach, which we call **Sparse Parameter Decomposition (SPD)**. SPD aims to decompose network parameters directly into a minimal set of simple subnetworks that are the shortest description of the mechanisms the network uses to compute its input-output behavior. We are cautiously optimistic about SPD and have spent the last 6 months developing the technique. We have focused on toy scenarios that are explicitly designed to test methodological choices. We have also red-teamed our approaches' theoretical assumptions. For example, our Circuits in Superposition [post](#) arose from discussions about the assumptions going into the method. We have found some encouraging results in these toy models (though challenges remain) and intend to publish an early write-up in 2025.

In our applied interpretability project, we aim to use **whitebox methods to detect deception in LLM agents**. We tested a variety of white-box techniques to test whether they can reliably detect different types of deception. We're particularly interested in whether these simple probes generalize to more realistic and agentic settings. Our preliminary results suggest that such probes could be used as one additional mechanism in a 'defense in depth' approach but are not reliable enough on their own. We will publish the paper in Q1 2025.

Our third interpretability project addresses field-level challenges. We were commissioned by a large philanthropic donor to develop a **proposal for a program** to support the field of mechanistic interpretability. To survey the landscape for this proposal, Lee Sharkey, our interpretability lead and CSO, has been writing and synthesizing a large review paper with many prominent mechanistic interpretability researchers that discusses the open research problems in the field.

In mechanistic interpretability, methodological progress has historically underpinned progress in downstream applications. However, we see many issues with current interpretability methods (some of which we helped develop). While there are many promising directions to pursue, most research in the field strongly correlates with directions pursued by frontier AI companies. We think there should be an organization outside frontier AI companies that focuses on improving and red-teaming interpretability methods and that is less correlated with industry research efforts. We would like to become that organization and believe we have the track record to do so. We are looking to work with a large funder to support this effort.

## Governance

We supported multiple jurisdictions with tailored policy advice throughout 2024 (see more [here](#)). As part of this, we provided our expertise via request advice, bilateral meetings and demos to international policymakers across institutions in south-east Asia, the European Union and the United States. For example, we drafted submissions to Requests for Information ('RFI') from

e.g. NIST, Singapore or the United Nations, participated in the AISI Network Meeting, and will be participating in the French AI Action Summit.

Complementary to that, we joined a select number of leading governance processes as expert contributors. In the EU, we are participants to the Code of Practice process. Internationally, we joined a working group defining 'red lines' run by SAIF, the US AISI Consortium, and standardisation efforts under JTC-21.

Targeted governance interventions and technical evaluations need a robust overarching system and comprehension to bear fruit. We therefore devote significant effort to raising public understanding of evaluations (see e.g. the [Evals Gap](#) and the [EU evals ecosystem](#)) and to strengthen the evaluation ecosystem itself.

Our focus in Q3 and Q4 2024 centered on the EU AI Act's implementation and publishing our internal thinking, e.g. on capability taxonomies and international information exchange. In 2025, we plan to focus on the governance of internal deployment and to pivot bandwidth to the UK and US governance ecosystem as their efforts concretise respectively.

## Operational Highlights

**Financials:** In 2024, we received philanthropic funding from numerous foundations and donors, as well as related business income from commercial contracts (e.g., governments and frontier AI companies). For each of the first three quarters of 2024, our expense spending is within 3% of our budget. Since our last update, we consistently maintained an average financial runway of nine months. Apollo aims to maintain a minimum of 6 months of runway at all times.

**People:** Over the past six months, we hired five full-time staff members, all of whom were our top-choice candidates. The seniority of our applicants has increased. Our new staff was able to contribute quickly to important research papers and run third-party Evals.

**Legal:** In Q4 2024, Apollo Research AI Foundation gained non-profit status in the U.S. Next quarter, we will spin out from our fiscal sponsor, enabling greater autonomy, cost savings, and tailored governance.

## Challenges

We **were able to successfully address most of the challenges** we have described in our 1-year update. We shared our work earlier and found it beneficial. We reduced the scope of our work without reducing their ambitiousness. We have regular calls with our advisors, which we found very helpful.

One challenge that we were unable to resolve is that leadership continues to have to spend a lot of time on **fundraising efforts**. This has meaningfully detracted from the time the leadership

has spent on technical work and management. To expand our most impactful work, we would like to hire more people and pay less uncompetitive salaries to attract and retain talent.

## Forward Look

- We want to make it **easier to understand, assess, and mitigate the risks related to deceptive alignment** and scheming for governments, AI developers, and civil society, for example, by evaluating frontier models for their scheming capabilities or demonstrating model organisms.
- We want to make it **easier to prepare for and address scheming-related risks**, e.g. by providing concrete evals & action plans like safety cases to governments and AI developers.
- We want to **establish white-box evals** as a useful and eventually required tool to assess safety and continue to improve interpretability methods.
- We want to **develop new interpretability methods** that supersede current flawed approaches.
- We would like to **diversify our interpretability projects** and grow our team. We think that the field of mechanistic interpretability would benefit from having a significant research pole outside of frontier AI labs.
- We want to help **establish and grow a healthy third-party evaluation ecosystem**.

We're very grateful for the support and trust we've been afforded so far, and we're excited to continue translating it into concrete outputs toward safer AI development and deployment.

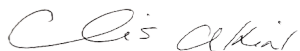
Kind regards,



*Marius Hobbahn, Chief Executive Officer*



*Lee Sharkey, Chief Strategy Officer*



*Chris Akin, Chief Operations Officer*



*Charlotte Stix, Head of AI Governance*

We are seeking additional funding. If you're interested, you may donate [here](#), or please feel free to [get in touch](#). We expect that increased funding would allow us to speed up and expand our most impactful research.