Schmidt Sciences

SAFE-AI: Safety Assurance through Fundamental Science in Emerging AI

Draft Research Agenda

September 2024

This document describes the program's current research agenda. We expect to regularly revisit and revise based on developments in AI and the progress of the AI safety research community.

Our research will focus on state-of-the-art LLMs, and on agent and multi-agent AI (i.e., AI that can act in the real world) and multi-modal settings (including speech, code, video, still images, and drawings).

We are interested in supporting scientific advances that can be broadly applied to safety criteria and testing methodologies for large classes of models and which would remain applicable even in the face of rapid technological advancement.

Note: "AI safety" is not a well-defined term and there is no universally agreed-upon set of safety properties that we would want to assert and prove about large language models or systems built with them. Generally speaking, the field of AI Safety deals with potential AI system behavior that might cause harm, and encompasses techniques to prevent this harm. The program will help the field make progress on these issues.

1) Basic research. Advancements in the science of AI safety

The rapid evolution and exploding scale of machine-learning-based AI systems present new challenges for their testing and evaluation. A number of AI safety-related benchmarks have emerged (e.g., Vidgen, et al., 2024; Bhatt et al., 2023) and a growing number of papers address frontier model performance against safety metrics (e.g., Phuong, et al., 2024). But the scientific basis for the creation of the benchmarks is rarely clear, and we are seeing benchmarks being surpassed rapidly as a result of frontier model improvement. So far, no clear way has emerged to know if an LLM is safe and trustworthy. This suggests two scientific themes for basic research into AI safety:

- How do we develop broadly-applicable and principled methodologies for creating benchmarks, based on a deeper understanding of the goals, architectures, training, and operation of AI systems?
- How can we make these methodologies resistant to rapid obsolescence from fast-evolving technology?

This research category is intended to catalyze the development of scientifically-based approaches to LLM testing and evaluation that will eventually result in robust evaluation methods that will be broadly applicable against all models (including those at frontier scale) and based on principles that allow them to remain useful as technology continues its rapid evolution. If we come to understand the principles that govern the relationship between data, training, learning mechanisms, fine-tuning, and inference, we may be able to design more robust and general ways of creating new benchmarks.

Research questions

A. Assurance: How can we develop assurances for specific model features, behaviors, and capabilities, especially in an agent or multi-agent environment where tangible effects—both good and bad—are possible? What measures and measurement methodologies can provide confidence for users of generative AI systems that their systems are safe to use? Assurance-related questions to be pursued include:

1. **Formal assurance**: What is the best way to think about assurance of AI systems? Are there approaches from statistics, theoretical computer science, or the science of metrology that can help us develop measurement approaches that can give us quantifiable levels of assurance? What formal confidence or probability measures are appropriate and how are they to be calculated and interpreted? Currently, we can guarantee simple properties of AI systems (Cohen et al., 2019), but complex safety properties currently remain out of reach. (Note: While formal verification is out of scope for this program for now, certified robustness is a special case of formal verification that holds promise and where we would like to support research.)

2. **Characteristics of high-quality benchmarks**: What role should benchmarks play in LLM evaluation? What formal characteristics of benchmarks can be articulated and how do they relate to predictive accuracy about future behavior of systems? How should thresholds be set for evaluated capabilities? Currently this is done only by intuition (Anthropic RSP, Yaghini et al., 2024).

3. **Contamination**: Generative AI systems are distinctive in basing their behavior on data collected from massive numbers of sources. As a result, prior tests and benchmarks for similar systems may very well be part of their training data, allowing them to "cheat" when tested. Improved performance can simply be due to memorization (Zhang et al.,

2024). Given that, how can test-set contamination be detected, including when semantically-related but not identical prompts can be used? How can testing be designed so that it probes general underlying capabilities rather than mere prior exposure to test-related information?

B. Generalizability: given the breadth of models that exist, even within the same model family, individual test results will be only modestly valuable unless they tell us about more than the single model under test. Given their scale and complexity, there are multiple dimensions along which it is challenging to generalize results of LLM testing.

4. **Transferability**: To what extent will conclusions drawn from testing open-weight models or smaller models be transferable to larger frontier models, or from current versions to future versions? Can empirical scaling laws be made scientifically rigorous (Sharma and Kaplan, 2022), and can they be used to make confident extrapolations on safety criteria? Can simple generalizable scaling models be inferred from multiple model families? Important scientific research has just begun to scratch the surface in this area (Ruan et al., 2024, Schaeffer et al., 2024).

5. **Multi-agent systems**: Interactions between multiple LLM agents carry the potential for emergent behaviors (Park et al., 2023), which may be hard to anticipate in single-agent evaluations. What do proofs about the safety of an individual agentic system tell us about the behavior of a collective of agents? We would welcome attempts to formalize competencies and failure modes of single- and multi-agent AI systems.

6. **Relative safety**: Are relative safety benchmarks, with measures such as win-rate and Elo ratings, useful? How robust are current measures and what new measures can be developed? Relative benchmarks are becoming more common, yet research on how to conduct these evaluations has only just begun (Boubdir et al., 2023). And pairwise comparisons tend to be fairly superficial, with one bit of information to express a comparison between two very complex objects.

C. Testing and evaluation frameworks: We believe generative AI systems are so large, varied, and rapidly changing that testing can be done adequately only if it is automated. In this context, several challenging problems arise:

7. **Agentic testing environments**: Given the increasing importance of agentic behavior, how can sandboxes or other testing environments be built that are realistically representative of the intended real-world execution environment for the system, but in which virtually any behavior can be safely tested? And how can we construct agent architectures/prompting strategies (e.g., can we measure/predict the delta between using only an LM and an LLM with an agent bolted on top)? Recent work has explored this question in LLM assistant settings (Ruan et al., 2023), but methods for building agent sandboxes remain largely

unexplored. This is especially challenging when the intended environment of use is the open internet.

8. **LLM-based scoring**: Some frameworks for testing, like the UK AI Safety Institute's Inspect, use LLM evaluators to assess target LLMs. Naive application of LLM-based-scoring can fall prey to biases and optimization pressure (Zheng et al., 2024). What is the science needed to understand how to assure the neutrality and accuracy of scoring done by LLMs on other LLMs' performance? Relatedly, how can neural network evaluators be made robust to optimization pressure (Gao et al., 2023)?

9. **Automated evaluations:** Is it possible to design and build a fully-automated ecosystem for testing LLMs? Such an ecosystem might need components for the various roles currently played by humans and testing technology in more conventional settings.

2) Applied research. Development of evaluations and evaluation environments

The goal of this category of work is to build high-quality benchmarks that address the challenges of existing benchmarks, and to use this more applied research to identify novel evaluation challenges, which would inform the theoretical work in Category 1.

We plan to fund research into evaluations of LLM capabilities that are currently a barrier to LLMs being more deeply integrated into important industries like education and healthcare. Examples of these capabilities include accuracy and reliability, hate speech and defamation, persuasion and deception, misinterpretation of intent, and privacy/confidentiality.

Out of scope research

At this point, this program will not pursue the following categories of research. We will revisit this list over the course of the program.

Red-teaming: We do not plan to fund research on human-based red-teaming of models, or the development of automated red-teaming approaches, as we believe this falls into the domain of "good and important engineering advances" rather than basic science.

Socio-technical research: We do not plan to fund research into non-technical aspects of AI safety, including ethics, policy, and governance research. These are hugely important issues and will be the focus of separate grantmaking efforts by us and others.

Formal verification: We do not plan to fund research into conventional formal verification methods. First, we are skeptical that it is currently possible for important and interesting safety properties of the largest LLMs to be extracted via conventional formal verification methods. Second, the UK ARIA's Safeguarded AI program plans to fund this direction of research.

Methods for improving AI safety: We do not plan to fund the development of new methods to specifically improve AI safety, such as improved methods of fine-tuning, activation editing, machine unlearning, and adversarial training. However, we expect some level of this work to flow naturally out of the research we fund, as the methods should result in general strategies. In addition, we expect these kinds of mitigations to be relevant in later stages of the program, and encourage proposers to keep in mind corrective actions they might consider to reduce safety risks uncovered by their tests.

Catastrophic risk: We do not plan to fund research into catastrophic risk, such as chemical, biological, radiological, and nuclear (CBRN) capabilities; cyber-security capabilities; self-proliferation; resource acquisition; and so on.