OpenAI

Careers

Research Scientist, Health AI

Safety Systems - San Francisco

Apply now ↗

About the Team

The Safety Systems team is dedicated to ensuring the safety, robustness, and reliability of AI models towards their deployment in the real world.

OpenAI's charter calls on us to ensure the benefits of AI are distributed widely. Our Health AI team is focused on enabling universal access to high-quality medical information. We work at the intersection of AI safety research and healthcare applications, aiming to create trustworthy AI models that can assist medical professionals and improve patient outcomes.

About the Role

We're seeking strong researchers who are passionate about advancing AI safety and improving global health outcomes. As a Research Scientist, you will contribute to the development of safe and effective AI models for healthcare applications. You will implement practical and general methods to improve the behavior, knowledge, and

reasoning of our models in these settings. This will require research into safety and alignment techniques that we aim to generalize towards safe and beneficial AGI.

This role is based in San Francisco, CA. We use a hybrid work model of 3 days in the office per week and offer relocation assistance to new employees.

In this role, you will:

- Design and apply practical and scalable methods to improve safety and reliability of our models, including RLHF, automated red teaming, scalable oversight, etc.
- Evaluate methods using health-related data, ensuring models provide accurate, reliable, and trustworthy information.
- Build reusable libraries for applying general alignment techniques to our models.
- Proactively understand the safety of our models and systems, identifying areas of risk.
- Work with cross-team stakeholders to integrate methods in core model training and launch safety improvements in OpenAI's products.

You might thrive in this role if you:

- Are excited about OpenAI's mission of ensuring AGI is universally beneficial and are aligned with OpenAI's charter.
- Demonstrate passion for AI safety and improving global health outcomes.
- Have 4+ years of experience with deep learning research and LLMs, especially practical alignment topics such as RLHF, automated red teaming, scalable oversight, etc.
- Hold a Ph.D. or other degree in computer science, AI, machine learning, or a related field.
- Stay goal-oriented instead of method-oriented, and are not afraid of unglamorous but high-value work when needed.
- Possess experience making practical model improvements for AI model deployment.

- Own problems end-to-end, and are willing to pick up whatever knowledge you're missing to get the job done.
- Are a team player who enjoys collaborative work environments.
- Bonus: possess experience in health-related AI research or deployments.

About OpenAI

OpenAI is an AI research and deployment company dedicated to ensuring that general-purpose artificial intelligence benefits all of humanity. We push the boundaries of the capabilities of AI systems and seek to safely deploy them to the world through our products. AI is an extremely powerful tool that must be created with safety and human needs at its core, and to achieve our mission, we must encompass and value the many different perspectives, voices, and experiences that form the full spectrum of humanity.

We are an equal opportunity employer and do not discriminate on the basis of race, religion, national origin, gender, sexual orientation, age, veteran status, disability or any other legally protected status.

OpenAI Affirmative Action and Equal Employment Opportunity Policy Statement

For US Based Candidates: Pursuant to the San Francisco Fair Chance Ordinance, we will consider qualified applicants with arrest and conviction records.

We are committed to providing reasonable accommodations to applicants with disabilities, and requests can be made via this link.

OpenAI Global Applicant Privacy Policy

At OpenAI, we believe artificial intelligence has the potential to help people solve immense global challenges, and we want the upside of AI to be widely shared. Join us in shaping the future of technology.

Compensation

$295K – $440K