OpenAI

Menu

Careers

National Security Threat Researcher

San Francisco, CA — Preparedness

Apply now ↗

About the team

Frontier AI models have the potential to benefit all of humanity, but also pose increasingly severe risks. To ensure that AI promotes positive change, we have dedicated a team to help us best prepare for the development of increasingly capable frontier AI models. This team, Preparedness, reports directly to our CTO and is tasked with identifying, tracking, and preparing for catastrophic risks related to frontier AI models.

Specifically, the mission of the Preparedness team is to:

1. Closely monitor and predict the evolving capabilities of frontier AI systems, with an eye towards misuse risks whose impact could be catastrophic (not necessarily existential) to our society; and
2. Ensure we have concrete procedures, infrastructure and partnerships to mitigate these risks and, more broadly, to safely handle the development of powerful AI systems.

Our team will tightly connect capability assessment, evaluations, and internal red teaming for frontier models, as well as overall coordination on AGI preparedness. The team's core goal is to ensure that we have the infrastructure needed for the safety of highly-capable AI systems—from the models we develop in the near future to those with AGI-level capabilities.

About you

We are looking to hire exceptional talent from diverse technical backgrounds (e.g., cybersecurity, CBRN-related expertise, national security/public safety) that can push the boundaries of our frontier

models. Specifically, we are looking for those that will help us shape our empirical grasp of the whole spectrum of AI safety concerns and will own individual threads within this endeavor end-to-end.

In this role, you will:

- Use your domain expertise to build our understanding of national-security-related AI safety risks
- Design (and then continuously refine) evaluations of frontier AI models that assess the extent of these risks
- Contribute to the refinement of risk management and the overall development of "best practice" guidelines for AI safety evaluations

We expect you to have:

- Hands-on experience with national security threat prevention, preferably in cybersecurity
- A deep interest in building understanding of the underpinnings of AI safety
- Familiarity with software engineering
- Ability to think outside the box and have a robust “red-teaming mindset”
- Ability to operate effectively in a dynamic and extremely fast-paced research environment as well as scope and deliver projects end-to-end

It would be great if you also have:

- Experience in ML research engineering, ML observability and monitoring, creating large language model-enabled applications, or another technical domain applicable to AI risk
- A good understanding of the (nuances of) societal aspects of AI deployment
- An ability to work cross-functionally
- Excellent communication skills

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.

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.

Annual Salary Range

$200K – $370K USD

Apply now ↗

Social
Twitter
YouTube
GitHub
SoundCloud
LinkedIn

Back to top