Evals Software Engineer

Hiring organization
Apollo Research

Job Location
London

Base Salary
$ 60000 - $ 110000

Date posted
June 17, 2024

Apply Now

Description

Applications deadline: The final date for submissions is 15 July 2024. However, we review applications on a rolling basis and encourage early submissions.About Apollo ResearchThe capabilities of current AI systems are evolving at a rapid pace. This provides us with many great opportunities but also brings challenges, including deliberate misuse or the deployment of sophisticated yet misaligned models. At Apollo Research, we are especially concerned with deceptive alignment, i.e. where a model appears aligned but is, in fact, misaligned and evades human oversight. Our approach involves conducting fundamental research on interpretability and behavioral model evaluations, which we then use to audit real-world models. Ultimately, our goal is to leverage interpretability tools for model evaluations, as we believe that examining model internals in combination with behavioral evaluations offers stronger safety assurances compared to behavioral evaluations alone.In our evaluations, we focus on LM agents, i.e. LLMs with agentic scaffolding similar to AutoGPT or SWE agent. We also fine-tune models to study their generalization capabilities and elicit their dangerous potential within a safe, controlled environment (see our security policies). At Apollo, we aim for a culture that emphasizes truth-seeking, being goal-oriented, giving and receiving constructive feedback, and being friendly and helpful. If youâ??re interested in more details about what itâ??s like working at Apollo, you can find more information here.nnThe Role. The evals team focuses on the following efforts:- Conceptual work on safety cases for scheming.- Building evals for deceptive alignment-related properties such as situational awareness or deceptive reasoning. – Running evals on frontier models and publishing the results either to the general public or a target audience such as AI developers or governments.- Building model organisms and demos of behavior related to deceptive alignment, e.g. how goal-directedness influences scheming.- Build a high-quality software stack that supports all of these efforts. Weâ??re looking for a software engineer who is excited to build software for these and similar efforts. We welcome applicants of all ethnicities, genders, sexes, ages, abilities, religions, sexual orientations, regardless of pregnancy or maternity, marital status, or gender reassignment.We are looking for the following characteristics in a strong candidate:Responsibilities- Maintain and extend our internal library for building and running language model evaluations.- Work closely with researchers to understand what challenges they face. Rapidly prototype, iterate on and ship useful features to increase their productivity.- Collaboratively iterate on the vision and priorities for the internal software stack.- Advocate for good software design practices and the general health of the codebase.- Keep up-to-date with the latest approaches to implementing evals frameworks and LLM scaffolding (e.g. open-source projects and research papers). -If interested, you can also run well-scoped research projects yourself, e.g. build an evaluation or do capability elicitation for an existing evaluation.Required Skills: At least 2 years of FTE-equivalent experience in software development with Python.Strong candidates may also have: Note that we welcome and value applications with different backgrounds. We encourage you to apply even if none of the following apply to you.- Experience in rapidly iterating on software products in close collaboration with users. For example, you have built software tools for internal or external users for 1+ years. – Strong communication skills and user/researcher empathy to understand and address their needs.- You enjoy close in-person collaboration and pair programming.- Experience building organization-internal tools. For example, you have led the efforts of an organization-internal tool for at least one 6-month project.Representative projects:- Design and implement a library feature that enables LLM agents to execute code in a sandboxed environment. – Develop a feature that lets us run a suite of evaluations

at the press of a button and saves results to a database in an easily accessible format. – Figure out how we can run evals presented in different formats by different organizations within our existing framework. – Extend our evals framework to support calling open-source LLMs in addition to the existing support for API LLMs.- Figure out what abstractions and conventions we can implement to reduce the amount of boilerplate and code duplication in our eval implementations.- Optimize the CI pipeline to reduce execution time and eliminate flaky tests.We want to emphasize that people who feel they donâ??t fulfill all of these characteristics but think they would be a good fit for the position nonetheless are strongly encouraged to apply. We believe that excellent candidates can come from a variety of backgrounds and are excited to give you opportunities to shine. About the Evals Team:The current evals team consists of Mikita Balesni, JÃ©rÃ©my Scheurer, Alex Meinke, and Rusheb Shah. Marius Hobbhahn currently leads the evals team. Weâ??re a small and tight-knit team, so you will likely interact with all members of the evals team on a regular basis. You will mostly work with the evals team, but you will likely sometimes interact with the interpretability team, e.g. for white-box evaluations, and with the governance team to translate technical knowledge into concrete recommendations. You can find our full team here.Logistics:- Start Date: Target of September/October 2024- Time Allocation: Full-time- Location: Our office is in London, and the building is shared with the London Initiative for Safe AI (LISA) offices. This is an in-person role. In rare situations, we may consider partially remote arrangements on a case-by-case basis.- Work Visas: We will sponsor UK visas for people who currently donâ??t have UK work permission.Applications deadline: The final date for submissions is July 15th, 2024. However, we review applications on a rolling basis and encourage early submissions.About the interview process: Our multi-stage process includes a screening interview, a take-home test (approx. 2 hours), 3 technical interviews, and a final interview with Marius (CEO). The technical interviews will be closely related to tasks the candidate would do on the job. There are no leetcode-style general coding interviews. If you want to prepare for the interviews, we suggest working on hands-on LLM evals projects (e.g. as suggested in our starter guide or get more familiar with public evals software like UK AISIâ??s Inspect). Benefits:- Private Medical Insurance- Flexible work hours and schedule- Unlimited vacationUnlimited sick leave- Lunch, dinner, and snacks provided for all employees on work days- Paid work trips, including staff retreats, business trips, and relevant conferences- A yearly $1,000 professional development budgetPlease mention the word **DEFEATED** and tag RMzUuMjM1LjEwNy40Mg== when applying to show you read the job post completely (#RMzUuMjM1LjEwNy40Mg==). This is a beta feature to avoid spam applicants. Companies can search these words to find applicants that read this and see they’re human.

Contacts

Job listing via RemoteOK.com