# **Research Engineer, Interpretability**

[(View all jobs)](https://boards.greenhouse.io/anthropic)

Remote-Friendly (Travel-Required) | San Francisco, CA | Seattle, WA | New York City, NY

## **About Anthropic**

Anthropic’s mission is to create reliable, interpretable, and steerable AI systems. We want AI to be safe and beneficial for our users and for society as a whole. Our team is a quickly growing group of committed researchers, engineers, policy experts, and business leaders working together to build beneficial AI systems.

## **About the role:**

When you see what modern language models are capable of, do you wonder, "How do these things work? How can we trust them?"

The Interpretability team’s mission is to reverse engineer how trained models work. We believe that a mechanistic understanding is the most robust way to make advanced systems safe. We’re looking for researchers and engineers to join our efforts.

People mean many different things by "interpretability". We're focused on mechanistic interpretability, which aims to discover how neural network parameters map to meaningful algorithms. Some useful analogies might be to think of us as trying to do "biology" or "neuroscience" of neural networks, or as treating neural networks as binary computer programs we're trying to "reverse engineer".

We [recently showed](https://www.anthropic.com/news/mapping-mind-language-model) that we could extract millions of meaningful features from Anthropic’s production Claude 3.0 Sonnet model, along with an initial demonstration of how we can use these features to change the model’s behavior by creating “[Golden Gate Claude](https://www.anthropic.com/news/golden-gate-claude)”. Achieving these results required a large engineering effort including optimizing sparse autoencoders (SAEs) across many GPUs, and building tools to visualize millions of features. Work like this is central to our roadmap of using mechanistic interpretability to improve the safety of LLMs like Claude.

A few places to learn more about our work and team are this [introduction to Interpretability](https://www.youtube.com/watch?v=TxhhMTOTMDg) from our research lead, [Chris Olah](https://colah.github.io/about.html); a [discussion of our work](https://open.spotify.com/episode/5UF79Uu94ia0fwC32a89LU) on the [Hard Fork podcast](https://www.nytimes.com/column/hard-fork) produced by the New York Times, and this [blog post](https://www.anthropic.com/research/engineering-challenges-interpretability) (and accompanying video) sharing more about some of the engineering challenges we’d had to solve to get these results. Some of our team's notable publications include [Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet](https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html), [Towards Monosemanticity: Decomposing Language Models With Dictionary Learning](https://transformer-circuits.pub/2023/monosemantic-features/index.html), [A Mathematical Framework for Transformer Circuits](https://transformer-circuits.pub/2021/framework/index.html), [In-context Learning and Induction Heads](https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html), and [Toy Models of Superposition](https://transformer-circuits.pub/2022/toy_model/index.html). This work builds on ideas from members' work prior to Anthropic such as the [original circuits thread](https://distill.pub/2020/circuits/), [Multimodal Neurons](https://distill.pub/2021/multimodal-neurons/), [Activation Atlases](https://distill.pub/2019/activation-atlas/), and [Building Blocks](https://distill.pub/2018/building-blocks/).

We collaborate with teams across Anthropic, such as Alignment Science and Societal Impacts to use our work to make Anthropic’s models safer.

## **Responsibilities:**

* Implement and analyze research experiments, both quickly in toy scenarios and at scale in large models
* Set up and optimize research workflows to run efficiently and reliably at large scale
* Build tools and abstractions to support rapid pace of research experimentation
* Develop and improve tools and infrastructure to support other teams in using Interpretability’s work to improve model safety

## **You may be a good fit if you:**

* Have 5-10+ years of experience building software
* Are highly proficient in at least one programming language (e.g., Python, Rust, Go, Java) and productive with python
* Have some experience contributing to empirical AI research projects
* Have a strong ability to prioritize and direct effort toward the most impactful work and are comfortable operating with ambiguity and questioning assumptions.
* Prefer fast-moving collaborative projects to extensive solo efforts
* Want to learn more about machine learning research and its applications and collaborate closely with researchers
* Care about the societal impacts and ethics of your work

## **Strong candidates may also have experience with:**

* Designing a code base so that anyone can quickly code experiments, launch them, and analyze their results without hitting bugs
* Optimizing the performance of large-scale distributed systems
* Collaborating closely with researchers
* Language modeling with transformers
* GPUs or Pytorch

## **Representative Projects:**

* Building [Garcon](https://transformer-circuits.pub/2021/garcon/index.html), a tool that allows researchers to easily access LLMs internals from a jupyter notebook
* Setting up and optimizing a pipeline to efficiently collect petabytes of transformer activations and shuffle them.
* Profiling and optimizing ML training, including parallelizing to many GPUs
* Make launching ML experiments and manipulating+analyzing the results fast and easy
* Creating an interactive visualization of attention between tokens in a language model

The expected salary range for this position is:

Annual Salary:

$315,000—$560,000 USD

## **Logistics**

Education requirements: We require at least a Bachelor's degree in a related field or equivalent experience.  
  
Location-based hybrid policy: Currently, we expect all staff to be in one of our offices at least 25% of the time. However, some roles may require more time in our offices.

Visa sponsorship: We do sponsor visas! However, we aren't able to successfully sponsor visas for every role and every candidate. But if we make you an offer, we will make every reasonable effort to get you a visa, and we retain an immigration lawyer to help with this.

We encourage you to apply even if you do not believe you meet every single qualification. Not all strong candidates will meet every single qualification as listed. Research shows that people who identify as being from underrepresented groups are more prone to experiencing imposter syndrome and doubting the strength of their candidacy, so we urge you not to exclude yourself prematurely and to submit an application if you're interested in this work. We think AI systems like the ones we're building have enormous social and ethical implications. We think this makes representation even more important, and we strive to include a range of diverse perspectives on our team.

## **How we're different**

We believe that the highest-impact AI research will be big science. At Anthropic we work as a single cohesive team on just a few large-scale research efforts. And we value impact — advancing our long-term goals of steerable, trustworthy AI — rather than work on smaller and more specific puzzles. We view AI research as an empirical science, which has as much in common with physics and biology as with traditional efforts in computer science. We're an extremely collaborative group, and we host frequent research discussions to ensure that we are pursuing the highest-impact work at any given time. As such, we greatly value communication skills.

The easiest way to understand our research directions is to read our recent research. This research continues many of the directions our team worked on prior to Anthropic, including: GPT-3, Circuit-Based Interpretability, Multimodal Neurons, Scaling Laws, AI & Compute, Concrete Problems in AI Safety, and Learning from Human Preferences.

## **Come work with us!**

Anthropic is a public benefit corporation headquartered in San Francisco. We offer competitive compensation and benefits, optional equity donation matching, generous vacation and parental leave, flexible working hours, and a lovely office space in which to collaborate with colleagues.