

[< Back to jobs](#)

AI Research Engineer (LLM Optimization)

PALO ALTO, CA

Apply

Chai is one of the fastest-growing, generative AI startups in Silicon Valley. YouTube but for LLM's - we have over 1 million active users.

Who we are looking for:

We need a relentless engineer with 3+ years of experience overseeing and being responsible for optimizing our LLMs. Ensuring they are performant, scalable, and cost-efficient. You will work alongside equally talented and driven teammates implementing cutting-edge AI inference engines. We need someone who is reliable and has high standards.

Here's why we might not be the right fit for you:

- We work hard and have a high-velocity environment with lots of growth opportunities.
- We value exceptional performance and continuous improvement. We believe that if you aren't constantly learning, you aren't growing.
- You will be responsible and accountable for making high-impact decisions that determine Chai's future

Here are the top 2 reasons why you should join us:

- Exponential growth. 1 Million MAU. Join the team that gets us to 100 million MAU
- Craftsmanship. Build something beautiful

Requirements:

- Familiar with vLLM, quantization, and current techniques of LLM optimization
- 3+ years of experience in software engineering
- Bachelor or Master degree from a leading academic institution

Here is our tech stack:

- Front end: Python, Flutter, Dart
- Back end: Python, GCP, Redis, Kubernetes

Process:

Exceptionally fast, application to offer within 7 days

1. Apply here

- 2. First round video interview, system design interview, then onsite
- 3. Reference checks, negotiation, and offer


Pay range

\$250,000 - \$350,000 USD

Apply for this job

* indicates a required field

First Name *



Last Name *

Email *

Phone

Location (City) *

[Locate me](#)

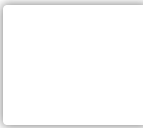
Resume/CV *

[Attach](#), [Dropbox](#), [Google Drive](#), or [enter manually](#)

Accepted file types: pdf, doc, docx, txt, rtf

LinkedIn Profile

Website



Submit application

Powered by [greenhouse](#)

Read our [Privacy Policy](#).