Full-Time Job

Principal Engineer, GenAI Blocks, Core ML - Mountain View, San Francisco or Sunnyvale

About this role

Full-Time Job (Level 8)

Location: Mountain View, San Francisco, Sunnyvale

Job family: Software Engineering

Job subfamily: Software Engineering

Product Area: TI Development

Updated Sep 5 | 348 views

Team members

Meenu Gaba
Mountain View
Sr. Director of Engineering, Core ML (hiring...

Kevin Hicar
New York
Executive Recruiter, Principal/Distinguished...

Description

Google is a leading ML/AI company. Google’s Core ML team aims to build the state-of-the-art AI platforms, services, and tools behind Google’s AI research and AI-powered products.

We are owners of the Google AI framework (e.g., TensorFlow, JAX), AI performance and efficiency, AI training and inference platforms, AI compilers (e.g. XLA), AI software/hardware co-design (e.g. TPU), and AI developer experience (e.g., Model Hub, Eval Hub). We are the driver to launch LMs (large models, like Gemini) from research to production and enable all Google products to adopt LMs to improve end-user experiences. The Core ML team is responsible for planning and optimizing ML capacity for all Google products and driving performance and efficiency work to maximize ROI for our AI investments. We look across Google’s AI efforts to build central solutions, break down technical barriers, and strengthen existing systems.

As the Core team, we have a mandate and a unique opportunity to impact important technical decisions across the company. As a Principal Engineer, you'll drive the technical strategy for orchestration of actions on top of GenAI models (and other AI models) that serve in our production serving system. This team delivers components for pre-/post-processing, RAI and safety, recitation, and sequencing of actions in our GenAI products from single model request calls to agent level orchestration that are available to all of Google. ***This system is in a critical path for all our GenAI launches.***

Minimum Qualifications:

- Bachelor's degree in Computer Science, Mathematics, other relevant Engineering field, or equivalent practical experience.
- 15 years of experience as a Software Engineering leader in ML Infrastructure, ML or AI for products, or related fields.
- Experience with machine learning frameworks and Google serving systems or indexing systems, and with developing technical strategy to support growth of the team, function, and customers.
- Experience with large-scale machine learning systems.

Preferred Qualifications:

- 15 years of experience effectively leading, growing, and scaling technical teams.
- Experience motivating others at all levels by creating a vision.
- Experience building ML infrastructure or products with heavy use of AI/ML and large groups of stakeholders or users.
- Experience in large language model, media generation, or other generative AI tuning and optimization techniques.
- Experience with working with stakeholders to understand their needs and translate them into technical requirements.
- Outstanding communication skills tailored to both technical and non-technical audiences.

Responsibilities:

- Define the long-term strategic vision for LM Root and Recitation technology and road map. Establish technical direction, goals, and development priorities including rethinking the strategy to meet the needs of the future.
- Forge strong partnerships with Research, Product, Hardware, and Core ML teams. Drive alignment and effective collaboration to ensure optimal solutions for Google's custom ML strategies.

Express Interest

Use the field below to send a message to this role's contact person. Your message is private: only that contact and the staffing team can see it. If you're in Tech, the hiring manager may request your GRAD from People Ops (see go/sharemygrad).

If you have questions about the role, but aren't yet ready to express interest, email the hiring manager listed to the left. If you're transferring into Google from another Bet, check out go/bet-mobility-faq first.

For more information regarding short term assignments, including eligibility, please visit go/transfers.

Have non-role specific questions about the Leadership Transfer Process at Google? Thinking about a Transfer with Level change? Curious about ways to optimize your application? Sign up for a confidential, 1:1 mobility session at go/leadershipmobility-oh or review curated resources at go/leadershipmobility .

To:

Kevin Hicar
New York
Executive Recruiter, Principal/Distinguish...

Transfer message*

EXPRESS INTEREST

- Deliver key infrastructure in GenAI applications for products and APIs that provides both essential functionality for high quality results, helps unify our stacks across Google, and supports safety and legal requirements at serving time.
- Apply your deep expertise in Serving and Indexing systems and machine learning to optimize system performance, reduce software overheads, and seamlessly support new hardware. Advocate innovation and the application of new technologies.

Please note:

This role is eligible for a transfer with a level change (TLC). To learn about Googler eligibility and process requirements for applying and potentially transferring into this role from one level below, see go/twlc and general transfer criteria.

- When submitting your application, please email your last two GRAD reviews (PDF) to Kevin Michael Hicar.
- We will review all applications and reach out as there is interest to move forward.
- Please do not reach out directly to the hiring manager for a coffee or info chat, as Google has moved away from pre-interview conversations to help remove bias, one way or another, from the interview and qualification process.

The US base salary range for this full-time position is $294,000-$414,000 + bonus + equity + benefits. Transfer compensation is determined algorithmically and is non-negotiable. Your recruiter will share more about the specific salary for your targeted location during the hiring process. Learn more about how a transfer may affect your compensation package, how location changes affect compensation, and about benefits at Google at go/benefits.

Skills:

- Generative AI
- Large Language Model
- Machine Learning Infrastructure
- TensorFlow Serving
- Indexing
- Machine Learning
- Software Engineering
- Google Infrastructure

○ My skills ○ Other skills