QCon SAN FRANCISCO by InfoQ

AI & ML at QCon San Francisco 2026

Sessions

Running Software Factory Efficiently at Uber Scale: Holding Agent Spend Steady While Requests Grew 9X

Monday Nov 16

At Uber, more than 70% of pull requests are attributed to coding agents. Engineers use more than 5K agent skills, and all MCP traffic goes through one gateway with more than 1,000 servers.

Uday Kiran Medisetty

VP, Distinguished Engineer @Uber

How Airbnb Guardrailed Its AI Customer Support Agent

Monday Nov 16

When Airbnb launched its AI customer support assistant, the feature made the news — the guardrails built to make it safe to ship did not. Yet an AI agent that talks to millions of customers, remembers a conversation across turns, and can trigger account actions is a

Weiping Peng

Distinguished Engineer @Airbnb, Previously Principal Architect @Salesforce for AI platform

---

Adapt or Drift: Resilience Engineering When AI Moves the Operating Point

Monday Nov 16

Resilient systems do not stay resilient by standing still. They survive by adapting. But adaptation comes with risk: under sustained pressure, organizations and systems can slowly drift toward failure while still appearing to operate normally.

Andrew Hatch

Engineering Leader and SRE Manager @Cisco ThousandEyes, With 25+ Years Building Software, Operations, SRE, and Platform Teams Across Australia, India, and the United States

QCon SAN FRANCISCO by InfoQ

Silent Agent Failures: Detect and Fix Them Before Your Customers Notice

Monday Nov 16

An agent run can receive a success verdict, complete every tool call without an error, and still fail the customer. It skips a necessary step, cites the wrong evidence, or delivers a confident answer that doesn't resolve the request.

Diane Lin

Staff Engineer @Datadog

How to Build a Real-Time Voice Agent

Monday Nov 16

A voice agent looks like a chatbot with a microphone.

Sahil Yadav

Software Engineer @Together AI - Building Real-Time Production Voice Agents

No New Backend: The API Layer Behind 800+ Agent-Built Apps at Duolingo

Monday Nov 16

At Duolingo, anyone (engineer or not) can describe an internal tool and get a working app back, deployed against real company systems. That's BuildingHub, and 350 Duolingo employees have created at least one app with it.

Sarah Deitke

Senior Software Engineer @Duolingo

Building Ultrafast LLM Inference at Scale: Disaggregated Inference across GPUs and Cerebras WSEs

Monday Nov 16

Ultrafast inference is a balancing act: users want tokens quickly, while providers need high aggregate throughput. The hardware choices that improve one can limit the other.

Nancy Zheng

VP of Engineering of AI Inference @Cerebras

Generation Got Cheap, Trust Didn't: Scaling Agentic Changes from Prompt to Prod at Roblox

Monday Nov 16

As AI agents write more of the code Roblox engineers push, the hard part is no longer generating changes; it's knowing which ones to trust.

Kushan Mehta

Principal Software Engineer @Roblox, Tech Lead for AI-Native Engineering

Jay Moorthi

Senior Engineering Manager, Diagnostics @Roblox

Agentic Trust: Securing AI Agents in Production

Monday Nov 16

This session will cover practical measures that you can take now to run agents more securely in production. This will not cover topics like model training and alignment; rather it will focus on deploying an existing agent as safely as possible.

Amanda Walker

Sr. Director, Engineering - Privacy, Safety and Security @Google

---

Into the Labyrinth: What Distributed Systems Have Taught Us About Guardrails

Monday Nov 16

AI has changed everything about software engineering...or has it? We've spent the past few decades learning how to build web-scale applications: complex, distributed systems that are too complex to fit into one reviewer's brain, but too critical to just YOLO into production.

Akshay Shah
Field CTO @Antithesis

It's Never the Thing That Broke: Learning From Incidents at (AI) Speed

Monday Nov 16

When something breaks, our first instinct is to focus on the thing that failed. But in complex systems, the failure is usually the result of many things going right, changing, and interacting in ways we didn't anticipate.

Vanessa Huerta Granda
Resilience Engineering Manager @Enova, Co-Author of the Howie Guide on Post Incident Analysis, Board Member for the Resilience in Software Foundation

---

The Schema Was Never Enough: The World in Protocols

Monday Nov 16

Every reliable process in the world is a protocol: auctions, double-entry bookkeeping, brokerage account transfers, or APIs. Protocols reduce ambiguity but the world remains messy.

Ryan Brown
Building Infrastructure @Render, Previously Principal @Crunchyroll, @Red Hat, and @Vendia

---

Skills, Memory, or Fine-Tuning? The Engineering Loop Behind Production Agents

Tuesday Nov 17

After deploying and improving hundreds of production agents, each handling millions of requests, we've seen the same patterns emerge in what works (and what doesn't) when trying to make agents better.

Abhinav Sinha
CEO @Lucidic AI, Previously @Stanford AI Lab, @Citadel and Susquehanna International Group, and @Apple

QCon
SAN FRANCISCO by InfoQ

Inside Asimov: BlackRock's Platform Architecture for Applied AI at Scale

Tuesday Nov 17

BlackRock's Portfolio Management Group is transforming how investment teams conduct research by combining deep investment expertise with shared AI capabilities.

Dan Wolf

Head of Platform Engineering for PMG Tech @BlackRock

Sri Annapurna Bhupatiraju

Director and Principal Engineer @BlackRock, One of the Founding Members of BlackRock AI Labs

Use AI to Multiply Your Team as a Staff Engineer

Tuesday Nov 17

Most "AI for engineers" content is about typing faster. That's IC-level work. Staff engineers have a different question to answer. When everyone has AI, what does a 10x engineer look like, and where does a staff engineer's leverage come from now?

Jordan Cutler

Senior Staff Engineer and Web Tech Lead @Pinterest, Author of High Growth Engineer Newsletter

Building a Data Platform with AI

Tuesday Nov 17

Data platform sits at the intersection of infrastructure, governance, operations, and user experience.

Mouli Mukherjee

Engineering Manager @OpenAI

The Revenge of the Data Scientist: Why Reliable AI Needs Evals, Traces, and Metrics

Tuesday Nov 17

Most teams can now ship an AI prototype by calling a foundation-model API. The hard part is knowing whether that system works when real users, messy data, and business consequences arrive.

Hamel Husain
Machine Learning Engineer, 20+ Years in Applied AI, Machine Learning, and Data Science

---

Lessons from Building a $100M Product in Six Weeks at OpenAI

Tuesday Nov 17

Engineering is shifting from synchronous, line-by-line implementation to an asynchronous model in which agents execute, verify, and retry while humans own intent, architecture, constraints, and release decisions.

Brian Yang
Member of Technical Staff @OpenAI

When Your Users Are Agents: Lessons from a Distributed Postgres Platform

Tuesday Nov 17

Distributed systems are built around assumptions about workload behavior: connections have reasonable lifetimes, retries eventually stop, traffic spikes have recognizable causes, and application code produces somewhat predictable query patterns.

Gwen Shapira
Co-Founder and CPO @Nile, Previously Engineering Leader @Confluent, PMC Member @Kafka, & Committer Apache Sqoop

---

Bridging Retrieval and Ranking: From LLM Embeddings and Graph Features to a GPU-Served DLRM

Tuesday Nov 17

The LinkedIn Feed retrieval layer selects, within 120 ms, the handful of posts each member sees from hundreds of millions of candidates. This split-second decision shapes what over a billion professionals read and engage with every day.

Sudarshan Srinivasa Ramanujam

Lead Engineer @LinkedIn - Working on Feed Ranking & Retrieval Modeling

---

Failing in Layers: Lessons from Rebuilding Netflix's ML Serving Platform

Tuesday Nov 17

Everyone talks about model quality. Fewer people talk about what happens when that model has to serve real traffic without falling over, or why the system ended up shaped the way it did in the first place.

Rajat Shah

Staff Software Engineer @Netflix

Wenbing Bai

Senior Software Engineer, AI Platform Team @Netflix

Adaptive Systems in Production: What Recommendation Systems Can Teach Us About Agents

Tuesday Nov 17

As organizations race to build AI agents, many teams are encountering challenges that feel new: evaluation uncertainty, feedback loops, behavioral drift, exploration versus exploitation, and maintaining user trust in systems that continuously adapt.

But these challenges are not new.

Mallika Rao

Senior Engineering Manager @Zocdoc, Previously @Netflix, @Twitter and @Walmart

---

How Netflix Retrofitted Its SDLC for AI Coding Agents

Wednesday Nov 18

Agentic tools amplify what's already in the codebase. The agent works through your code, your architecture decisions, and the documentation around them. It follows what it can read and guesses at the rest.

Mark Khuzam

Senior Software Engineer @Netflix, Previously Led the Consumer Web Platform Team @OpenTable

---

No More Small Systems, How AI May Bring Large System Problems to Everyone

Wednesday Nov 18

Once you figure out how to use AI to make your organization productive, you will soon find you have a whole new set of problems. While some of those problems will be novel thanks to LLMs, many of them are just the result of there being "more stuff".

Adam Bender

Technical Lead for Enterprise Engineering @Google

Efficient GPU Agents: Infra for Autonomous LLM Perf Optimization

Wednesday Nov 18

Parasail serves open-weight LLMs on distributed clusters of GPUs managed by Kubernetes. The typical performance optimization loop is to reproduce a problem or measure a regression, iterate on code or config, set up and run benchmarks, and decide whether the change is worth shipping.

Ben Segal

Software Engineer @Parasail

Meghana Madhyastha

Performance Engineer @Parasail, Previously Fundamental AI Research Lab @Meta

---

Observability: The Missing Control Plane for Agentic Engineering

Wednesday Nov 18

As AI agents move from generating code to testing, deploying, and operating it, the important question is no longer “How autonomous is the agent?” It’s “What can this agent safely decide without a human?”

Rachael Wonnacott

Global Head of Machine Learning and Applied AI Engineering @Capital Group, OpenUK Ambassador & Founding Member of FastFlowConf, Previously Head of Cloud Engineering @Fidelity International

QCon
SAN FRANCISCO by InfoQ

Product Engineering: When code costs nothing, what's next?

Wednesday Nov 18

Coding agents produce 741% more code, and only 30% more software ships as a result. The gap is human review: people top out around 400 lines an hour and agents are not slowing down. This talk is for engineering leaders trying to decide what to do about that.

Laurie Voss

Head of Developer Relations @Arize AI, Co-Founder of npm, Inc., Developer for 30+ Years

---

You Can't Read the Code Anymore - Verifying Software in the Age of AI

Wednesday Nov 18

AI writes most of the code now, so the hard part of engineering has moved from authoring to the part we're worse at - verifying. With the incentives all pushing toward shipping faster and faster - the rational move is to skim the diff, click through a little manual QA, and ship on vibes.

Omer van Kloeten

Principal Engineer @Forter, 25+ Years in Software Engineering, Now Building an AI-Agent Framework to Tame Legacy Systems and Reduce KTLO

Avoiding Prompt Debt & Other Context Failures

Wednesday Nov 18

As agents evolve, prompts get unruly. Edge cases are patched with one-line hotfixes, or with stacked instructions of escalating severity when an issue proves stubborn. Over time, those instructions pile up into a brittle artifact, overfit to an aging model.

Drew Breunig

CEO & Cofounder CMPND, Editor of O'Reilly's "Context Management Handbook", Writer on AI & Data at dbreunig.com, Previously Data Science & Product Lead PlaceIQ

QCon SAN FRANCISCO by InfoQ

Trainings

[SOLD OUT] Building AI Agents: A Practical Framework for Choosing the Right Agent Harness

Thursday Nov 19

Every agent framework lists memory, tools, context management and guardrails. The feature lists read the same.

Hien Luu
AI/ML Leader, Advisor, Speaker, and Author

AI-Assisted Coding Master Class: Build Your Mission Control

Thursday Nov 19

Most engineers aren't spending enough time setting up their environment to get the most out of their coding agents.

Sepehr Khosravi
Founder @ OneKit | Award Winning Faculty @ UC Berkeley | Prev. ML Coinbase, Tesla
Dev Bali
ML Systems PhD Student @ UC Berkeley
