Haiyuan Cao

Staff Software Engineer at Google · AI agent observability, evaluation and coding-agent tooling

Greater Seattle Area · linkedin.com/in/haiyuancao · github.com/haiyuan-eng-google · github.com/caohy1988 · caohy1988.github.io

Staff software engineer with ten years in applied machine learning and LLM systems. Tech lead for BigQuery Agent Analytics at Google: agent traces in BigQuery and the open-source tools to analyze and evaluate them. Previously Principal Software Engineer at Microsoft on GitHub Copilot and Azure OpenAI. Named inventor on 9 granted US patents.

EXPERIENCE

Google · Staff Software Engineer; Tech Lead, BigQuery Agent Analytics

Kirkland, WA · 2025 – Present

- Tech lead for BigQuery Agent Analytics, which records what AI agents do as traces in BigQuery so teams can analyze and evaluate them. Generally available since Google Cloud Next ’26.
- Lead development of its open-source SDK for trace analysis and agent evaluation, working with a team of AI agents under my direction; 21 releases published since April 2026.
- Extended tracing to coding agents with a Claude Code plugin and mod, so their sessions can be analyzed alongside agents built with Google’s Agent Development Kit (ADK); co-maintain an OpenTelemetry Collector exporter for BigQuery.
- Contribute to ADK itself, its Go agent analytics package and its documentation, and built an agent-native command-line tool for Google Cloud data services.

Microsoft · Principal Software Engineer

Bellevue, WA · 2017 – 2025

- **GitHub Copilot:** improved code understanding and predictive edits for Agent Mode and Next Edit Suggestions.
- **Azure OpenAI On Your Data:** tech lead for AI quality across retrieval-augmented generation (RAG) scenarios; built the evaluation pipelines, synthetic data generators and benchmark datasets used to assess and improve it.
- **Power Apps and Power BI:** built LLM-based natural-language-to-code translation into Power Fx and DAX.
- **Earlier:** machine-learning work on scheduling and workflow tools, the source of five granted US patents (2020 – 2023).

Before Microsoft: Researcher, Columbia University Medical Center (2016 – 2017) · Data Scientist Intern, Tapad (2016)

PATENTS

Named inventor on 9 granted US patents, with 2 further inventions pending, all assigned to Microsoft. The granted patents cover training-data generation for natural-language-to-code models, code generation options for data structures, a domain-integrated contextual response engine for AI systems, and machine-learning scheduling and workflow engine tools. Pending: content quality evaluation for retrieval-augmented generation (RAG) systems, as first-named inventor (US 2026/0099693), and function calling for multi-source data retrieval in generative AI systems (US 2025/0384064).

Granted: US 12,608,181 · 12,596,721 · 12,578,932 · 12,159,115 · 11,734,066 · 11,327,726 · 11,068,304 · 10,768,908 · 10,579,423

OPEN SOURCE AND WRITING

- Build in the open with a team of AI agents working under my direction: 850+ pull requests merged across two GitHub accounts as of Nov 10, 2026, many of them produced with that team.
- Publish the team’s playbook, a blog series on agent analytics and evaluation, and independent writing: Durable running agents, explained and Technical leadership in the AI agent era.

EDUCATION

Columbia University, M.S. in Data Science, 2015 – 2016

Fudan University, Ph.D. in Theoretical and Computational Physics, 2010 – 2015; B.S. in Physics, 2006 – 2010

10 journal papers, including Physical Review Letters and Nature Communications; first-author Web of Science Highly Cited Paper.

SKILLS

Languages: Python, Go, Rust, TypeScript, SQL · **Evaluation:** LLM and agent evaluation, LLM-as-judge, CI quality gates, RAG evaluation · **Agent infrastructure:** tracing (OpenTelemetry, BigQuery), ADK, MCP, agent skills and plugins