HAO HU

+1 (408) 593-4072 · felixhuhao@gmail.com · haohu.tech

SUMMARY

Senior backend engineer, now an AI-focused CTO — 10+ years building large-scale, high-availability distributed systems and data infrastructure, including at Google, Uber, and Microsoft.

Builds production agentic AI systems — RAG, conversational analytics, and MCP-backed multi-agent orchestration, with harness engineering across golden-set evaluations, deterministic guardrails, and human-in-the-loop safety.

Hands-on engineering leader — leads teams from architecture through delivery with spec-driven, review-gated AI-assisted workflows, quality gates, and automated validation.

TECHNICAL SKILLS

LLM & Agents: agent orchestration, agent governance, tool / function calling, MCP integration, multi-agent routing, context engineering, DeepAgents / LangGraph / LangChain, OpenAI / Claude / DeepSeek

RAG & Retrieval: RAG (multimodal), hybrid retrieval, embeddings, reranking, query optimization, Milvus, Qdrant

Evaluation & Safety: LangSmith, RAGAS, OpenEvals, LLM-as-judge, guardrails, human-in-the-loop, observability

Programming Languages: Java & Go **(Google readability)**, Python, SQL, C#, C++

Backend & Distributed Systems: microservices, API design, gRPC / Protobuf, REST, Thrift, Spring AI, FastAPI, SSE

Data Systems: data modeling, PostgreSQL, MySQL, Redis, BigQuery, Kafka, Flink, ClickHouse, MongoDB

Cloud & Platform: Google Cloud, AWS, Docker, Kubernetes, Linux, CI/CD

EXPERIENCE

Yuanbao Creative Technology **Feb 2025 – Present**

Chief Technology Officer *Wuhan, China*

- **Architected and shipped a unified AI layer for an ERP SaaS platform** across operations, analytics, and knowledge, leading a team of 4; established shared eval/safety practices: observability, golden sets, and regression suites.
- **Operations Copilot**: built a **DeepAgents**-based agent harness with delegated specialists (sales / inventory / order / purchasing), using role-scoped **MCP** tools, sandboxed execution, proactive monitoring, auditable tool traces, and a **Spring AI / Java 21** server with human-in-the-loop approval for risky writes (cryptographically bound, single-use).
- **Analytics Agent**: delivered self-service, multi-turn NL-to-SQL over the ERP **ClickHouse** warehouse, backed by a **Qdrant** semantic layer that cut prompt context by **~73%**; added OLAP intent routing, deterministic **SQLGlot** guardrails, read-only execution, bounded SQL repair, result-equivalence regression evaluations, and auto-charting.
- **Knowledge Assistant**: delivered **LangGraph**-orchestrated enterprise-document RAG with multi-source ingestion and **Milvus** hybrid retrieval; routed precise / balanced / broad strategies with query expansion and rerank, enforcing citation grounding, strict-evidence refusal, and RBAC; added LLM-judge/citation evaluations and per-query observability.

Independent **Mar 2023 – Jan 2025**

Independent Software Engineer & Consultant *Toronto, ON / Wuhan, China*

- **Built and deployed a cloud-native travel-guide SaaS** with QR-based POI audio guides, anonymous ratings, and playback analytics, using Next.js + Go on AWS (RDS Postgres, S3/CDN) and Vercel, Docker container deployment.
- **Developed RAG/agent prototypes**, building hands-on AI engineering depth into the CTO role.

Uber **May 2022 – Mar 2023**

Senior Software Engineer *Toronto, ON*

- **Built Uber's first CPG (consumer-packaged-goods) Ads feature — storefront banner ads**: as backend lead on a 3-person team, built a Go/gRPC + DocStore service between UberEats Storefront and Ads so advertisers could run budgeted item/brand campaigns across retail chains; **scaled to 10–30K QPS at ~140 ms p99** within a quarter.
- **Redesigned CPG Ads campaign data model**: migrated from restaurant-scoped item IDs to a GTIN-keyed multi-location model — enabling national campaigns, removing millions of duplicates, and forming CPG Ads foundation.
- **Launched store-menu sponsored items** by extending the CPG Ads backbone into a third-party ad integration: ranked sponsored items in storefront sub-menus, streamed impression/click events through a Kafka + Flink billing pipeline, and **held ~180 ms p95** with Prometheus / Grafana monitoring and dependency alerting.
- **Drove the staged production rollout of CPG Ads** and served as the Ads team's dev representative — aligning product, data science, sales, DevOps, partner teams, and the external ad provider.

Google **Feb 2018 – Feb 2021**

Software Engineer — Cloud SQL (Data Integration / Disaster Recovery) *Sunnyvale, CA*

- **Raised backup health from 99.92% to 99.97%** across a large fleet under a 99.95% SLA: owned and hardened the backup-verification pipeline (snapshot attach, DB boot, checksum on pooled GCE instances), resolving top failure classes including resource exhaustion, startup failures, and false-positive verification failures.
- **Enabled PITR-safe transaction-log retention**: built backup PD-snapshot cleanup that pruned nonessential general-log data, preserved recovery-critical logs, and allowed restore to any timestamp via nearest backup + log replay.
- **Redesigned backup retention**: replaced a fixed-count window that pruned oldest-first regardless of health with a policy that always kept the most recent healthy backup, guaranteeing each instance had a recoverable restore point.
- **Automated MySQL HA migration to Persistent Disk synchronous replication**: built a Java workflow converting legacy streaming master/replica pairs, spanning both control and data planes.
- **Built customizable CSV formatting across MySQL and PostgreSQL**: extended the backend REST API and redesigned import / export one-shot tasks in Go, enabling reliable export / import round-trips.
- **Shipped Cloud SQL PostgreSQL CSV import / export to GA**: hardened the injection-sensitive import / export path with on-demand, least-privilege database access and Cloud SQL IAM for Google Cloud Storage flows.
- **Upheld the 99.95% SLA as a data-plane on-call engineer** — resolved release-blocking and customer-escalated incidents across the Java backend and Go on-instance agents; mentored junior engineers.

Microsoft **Jan 2014 – Oct 2017**

Software Development Engineer *Mountain View, CA*

- **Co-built "People I Communicate With" (PICW) from the ground up** on a 3-person team: a foundational Exchange signal-collection component in C#, capturing frequent contacts and mail-interaction signals for Contact Safety and Exchange anti-spam, contributing to safety-level classification, junk-rule processing, and spam/phishing detection.
- **Took sole ownership of PICW**, expanding its signal surface and owning cross-team integration with Exchange anti-spam, Contact Safety, and People Relevance consumers; its signals became a core input to People Relevance, powering Graph People API suggestions, ranking, and lightweight CRM.
- **Built the monitoring backbone for People Relevance**: log dashboards, real-time alerts, and performance counters gave the team operational visibility into the service powering the Graph People API.
- **Built a distributed C# profiling tool that drove account selection for the Hotmail / Outlook.com-to-Exchange migration**: profiled feature usage across the 2.2B-account base via SQL scans, surfaced 40+ feature gaps, identified 400M+ migration-ready accounts, and re-profiled as gaps closed for successive migration waves.
- **Served on the Exchange on-call rotation** across Hotmail / Outlook services, triaging release-blocking and customer-escalated issues and shipping fixes and improvements.

Epic Systems **Jun 2012 – Nov 2013**

Software Developer *Madison, WI*

- **Built and extended EMC2**: Epic's internal software configuration management platform (VB.NET on InterSystems Caché) for version control, dependency management, and automated build / deployment — an internal CI/CD system.
- **Developed customer release and upgrade utilities** (console and web) that automated version checks and pushed updates over a secure communication framework, keeping health systems on current, supported versions.
- **Extended a Trial Dev Environment** for safe high-risk change integration and resolved build / deployment escalations.

HCR ManorCare **Jan 2011 – Apr 2012**

Data Warehouse Developer *Toledo, OH*

- **Built SSIS ETL pipelines feeding the enterprise data warehouse of a 500+ facility national healthcare provider**: extracted, validated, transformed, and loaded data from multiple sources into conformed fact/dimension schemas.
- **Developed SSRS reports and dashboards** (cross-tab, drill-down, parametric, cascaded) for finance/operations BI.
- **Created SQL Server objects** (stored procedures, indexed views, UDFs) powering report logic and ETL validation.

EDUCATION

Bowling Green State University Aug 2010 – May 2012

M.S. in Computer Science GPA 4.0/4.0 · Bowling Green, OH

Wuhan University Aug 2006 – Jun 2010

B.S. in Software Engineering GPA 3.5/4.0 · Wuhan, China