

ADAM WOLD

Staff Software Engineer | AI Agents & Evaluation

Golden Valley, MN • 320-224-8602 • adam.n.wold@gmail.com • linkedin.com/in/adam-wold-b708582a

PROFESSIONAL SUMMARY

I have 14 years of engineering behind me and spent the last one building AI agents and the evaluation infrastructure that tells a team whether they actually work. I shipped production agent workflows on Claude, typed tools, MCP servers, subagents and state carried across multi-step runs, then wrote the evaluation platform underneath them, sole author, so a model, prompt or pipeline change is measurable as a delta instead of an opinion. Before that I was a founding engineer twice at the same company and then tech lead for its team. All of it ran on regulated healthcare records, where audit trails, permissions and data residency were requirements rather than good intentions.

EXPERIENCE

Staff Software Engineer (Team Lead) | Cellares

Mar 2022 – Aug 2026

- Designed and shipped production agent workflows on Anthropic Claude, built on a TypeScript agent framework: typed tools, MCP servers, OpenAPI connections, specialist subagents, and state carried across multi-step runs against live operational data. Every action that writes stops at a human commit step, which is the decision that made the system deployable at all in a regulated setting.
- Built the company's AI evaluation platform from scratch, sole author, about 20,500 lines of Python: golden datasets from real production records, retrieval metrics plus LLM-as-judge scoring, and CI gates that block a release on failing evaluations. Three interchangeable systems-under-test sit behind one contract, so any engineer's prompt or pipeline change is measurable as a delta rather than an opinion.
- Designed the judge methodology rather than taking one off the shelf: binary criteria instead of Likert scales, rationale enforced before verdict through output field ordering, and a fail-closed gate that reports failure when it checked nothing instead of passing vacuously. Established the harness's noise floor and a minimum actionable effect size, so small movements stopped being read as progress.
- Took the worst failure mode, the model recommending that a genuine quality event be dismissed, from 3 occurrences to 0 across 201 cases, kept embeddings and reranking in-environment so regulated content never left it, and caught 5 platform defects and a 14% document-library gap before any user saw them.
- Ran a 10-experiment retrieval campaign against a 499-document corpus and raised accuracy from about 3-in-10 to 8-in-10 through section-aware chunking, embedding-input enrichment and document grouping. Documented 5 changes that measured worse and were dropped, including cross-encoder reranking, the one everyone assumes will work, which cost a net 21 cases.
- Exposed the product wiki as an MCP server so AI clients across the company reuse the same grounded context, then built the internal application on it that drafts work items and test cases against real API specifications, adopted by 11 colleagues, and cut cost 47 to 62% at the same quality through a benchmarked model and SDK migration.
- Largest contributor, 40% of all pull requests, on a new AI-assisted quality operations product in Next.js, React and Postgres, covering record intake with chain-of-custody hashing, a role-aware workspace, append-only audit tables and the surface where the model's recommendation reaches a reviewer. Ran it on Azure Container Apps with Key Vault, managed identity and private networking.
- Founding engineer twice at the same company, six years apart, and tech lead for the team, mentoring engineers through near-daily office hours and owning release scope planning and sign-off. On the protocol-authoring platform I wrote the second commit in the repository and finished as the top contributor of 28 engineers, and designed its real-time telemetry backbone as sole author.

Senior Software Engineer | Cellares

May 2020 – Mar 2022

- Joined as Cellares's 13th employee, before the software suite existed, as founding engineer on the manufacturing-execution web platform. Built Process Designer end to end in Angular and .NET, along with the authorization and permissions model and the 21 CFR Part 11 audit trail UI.

Associate Consultant → Consultant → Senior Consultant | Magenic Technologies

Feb 2012 – Jul 2019

- Embedded with enterprise clients across travel, entertainment, financial services and payments, owning a delivery workstream end to end each time, from requirements through architecture to production. 2014 Consultant of the Year.

ADDITIONAL EXPERIENCE

Senior Software Engineer, Exact Sciences — genomic testing and cancer-detection platforms (Aug 2019 – May 2020)

TECHNICAL SKILLS

Agents & LLM Systems: Agent orchestration on Anthropic Claude (typed tools, MCP servers, OpenAPI connections, subagents, skills), multi-step tool use, state management, human-in-the-loop controls, prompt and context engineering, structured outputs, RAG end to end, pgvector (HNSW, hybrid retrieval)

Evaluation & Guardrails: LLM-as-judge methodology, golden datasets built from production records, retrieval metrics (recall / MRR / NDCG), RAGAS, hallucination detection, CI quality gating, noise-floor and effect-size discipline, de-identification and data residency, append-only audit trails

Languages & Frameworks: Python (FastAPI, pytest), TypeScript and JavaScript, React, Next.js, Node.js, C# and .NET Core (14 yrs), Angular, SQL

Architecture & Platform: Ports and adapters, event-driven messaging (RabbitMQ, SignalR), REST / OpenAPI / gRPC service design, feature flags and kill switches, RBAC, OpenTelemetry

Cloud & DevOps: Azure (Container Apps, Key Vault, PostgreSQL Flexible Server, managed identity, VNet and private endpoints), Docker and Compose, Kubernetes (AKS, ArgoCD), Azure DevOps Pipelines, SonarQube

Domain: Regulated healthcare and life sciences, 21 CFR Part 11, review workflows, technical leadership and mentoring

EDUCATION

Bachelor of Science, Computer Science — University of Minnesota (Focus: Databases and Software Engineering)