Henrik Albihn, MS

AI engineer · Forward-deployed · Founder

Irvine, CA · (714) 397-4697 · h@henrikalbihn.com · henrikalbihn.com
github.com/henrikalbihn · linkedin.com/in/henrikalbihn · x.com/henrikalbihn · medium.com/@henrikalbihn · hf.co/henrikalbihn

SUMMARY

Owned the size-recommendation engine at **True Fit** — the product the company sells — sizing millions of shoppers daily across 30K+ brands including Nike, Target, and Walmart. Principal scientist on a manufacturing copilot running on Meta Quest 3 and Apple Vision Pro. Now AI engineer at **Strange Loop Labs**, forward-deployed into Fortune 500 and Big Four engagements: exec discovery through production handover.

Economics → data science → ML science → applied AI → forward-deployed engineering. At every stop I owned the system the product depended on. Current work: classical ML applied to the code-generation loop, and the open-source tooling underneath it — **ticket-rs.io**, **sgrep.sh**.

EXPERIENCE

AI Engineer · Strange Loop Labs

Feb 2025 → present · remote · US

Early engineer at a company building the Autonomous Firm: specialized AI agents that automate financial institutions' operations. Own the technical architecture and delivery for each engagement, exec discovery through production handover.

- Regulated document workflows for Fortune 500 and Big Four clients — extraction and validation over the unstructured formats these firms actually run on, deployed into SOC-2 Type 2 environments on AWS (Lambda · ECS · API Gateway · S3 · SQS) with observability and runbook handover.
- **Internal research: high-throughput agentic coding.** Classical ML applied to the code-generation loop rather than prompt engineering alone — ensemble methods, online learning, statistical quality control. The harness decomposes customer call transcripts into dependency-ordered tickets in ticket-rs and dispatches coding agents that inherit the exact context the FDE heard.
- Technical lead on customer C-suite readouts; shaped the hiring bar and the technical interview loop as the team scaled.

Principal AI Scientist · Stealth AR/VR Startup

Oct 2024 → Apr 2025 · remote · US

Owned the AI system behind the company's only product: a manufacturing copilot on **Meta Quest 3** and **Apple Vision Pro** giving real-time assembly guidance to production-floor workers. Reported to the CEO; two direct reports. Owned system architecture, technology selection, and technical go-to-market.

- Built YOLO-based real-time video pipelines and vision-LLM systems that infer assembly workflows from training video and live camera feed. That removed the per-customer setup cost that had been blocking every deployment.
- Real-time multimodal inference (video · audio · text · spatial / LiDAR) across a multi-node GPU cluster, with self-monitoring, canary rollout, and automatic rollback.

Machine Learning Scientist · True Fit

Mar 2022 → Aug 2024 · remote · US

Owned the size-recommendation engine — the product True Fit sells — sizing millions of shoppers daily across **30K+ brands**: Nike, Target, Walmart, Macy's, JCPenney, J.Crew, Dick's Sporting Goods, Urban Outfitters, Lululemon, Gap, Nordstrom. PyTorch training on True Fit's cross-retailer dataset; served on GCP Vertex.

- Rebuilt the A/B testing platform from parametric to non-parametric statistics. Its chronic false-positive rate had been driving roadmap decisions; fixing the measurement changed what the company built next.
- Re-architected the recommender from legacy regression to a **3-class deep-learning ensemble**, shipped through dozens of A/B-tested release cycles.
- Replaced years of regex and manual curation with LLMs (GPT · Llama · Mistral · Gemini) and BERT embeddings for product-attribute extraction, taxonomy classification, and reverse-image search.
- Worked directly with retailer data teams. Most of what shipped came out of those conversations, not out of internal roadmap planning.

Data Analyst, Institutional Research · Westcliff University

Mar 2021 → Feb 2022 · Irvine, CA

- Built university outcome-prediction models; automated Python / SQL ETL across every campus

information system.

VENTURES

Thesis: the harness around the model — context, search, data access — is where the reliability gap sits. All three tools came out of work I was doing, and all three are in my own daily loop.

ticket-rs.io · Founder · 2026 → present
AI-native issue tracking. Git-backed, local-first, no daemon. Treats the backlog as a dependency graph and runs **PageRank prioritization** and critical-path analysis over it to decide what to build next. MCP servers plus a Python SDK, so it works with any MCP-speaking agent; GitHub and Linear sync. MIT-licensed, public alpha.

sgrep.sh · Founder · 2026 → present
Semantic search for codebases agents write faster than people read them. Single static Rust binary. 8M-parameter **Model2Vec** embeddings on **Candle**, HF's Rust ML framework; sub-10ms search; hybrid semantic + BM25 via ripgrep. MIT-licensed.

SQLGenie · Founder · 2024 → 2026
Natural-language query engine across **23+ SQL dialects**. Invited by Techstars NYC to apply (Nov 2024); met with Surface Ventures.

WRITING & OPEN SOURCE

Vanishing Gradients — blog on production AI, harness engineering, and the economics of coding agents. Recurring argument: most agent failures are context failures, not model failures — the scaffolding is the product. Republished in *AI in Plain English* and *Level Up Coding*.

gliner-as-a-service (47 ★) · seven models on **Hugging Face** — NeuralPipe (SLERP/TIES merges), NeuralHermes (GGUF quantization), Qwen2.5-Sci.

EDUCATION

M.S., Data Science — California State University, Fullerton · *User-product recommendation via matrix factorization and SGD*
B.A., Economics — California State University, Fullerton · *Demand decomposition in the plant-based meat-substitute market*

SKILLS

AI — LLM fine-tuning (LoRA · SFT · DPO · RLHF) · model merging · quantization · RAG · evals · multi-agent orchestration · computer vision (YOLO · object detection) · multimodal (vision / audio / text / spatial) · recommender systems · NER

Systems — AWS (Lambda · ECS · API Gateway · S3 · SQS) · GCP Vertex · Docker · Kubernetes · Terraform · low-latency serving · GPU clusters · MLOps · MCP

Frameworks — PyTorch · Hugging Face Transformers · LangGraph · vLLM · pgvector / FAISS

Languages — Python · TypeScript · Rust · Go · SQL · Bash