Henrik Albihn, MS

Serial founder · AI engineer · Forward-deployed

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

Shipped **True Fit's** size-recommendation engine — the company's flagship product — to millions of real shoppers daily across **30K+ brands** including Nike, Target, and Walmart. Principal scientist on a real-time **manufacturing copilot** deployed on **Meta Quest 3** and **Apple Vision Pro** at a stealth AR/VR startup. AI Engineer at **Strange Loop Labs**, forward-deployed into **Fortune 500** and **Big Four** engagements. Founder of **ticket-rs.io** (AI-native issue tracking; public alpha), **sgrep.sh** (semantic code search), and **SQLGenie** (NL-to-SQL, 23+ dialects; TechStars / VC / acquirer inbound in the first 90 days).

Arc: **economics** (causal inference, demand modeling) → **data science** → **machine-learning science** → **applied AI** → **forward-deployed AI engineering** — each transition sharpened the same edge: pattern recognition across markets and systems. Current research: applying **classical ML techniques to agentic coding** — pushing the boundaries of what's possible in code generation beyond prompt engineering alone.

VENTURES

Thesis: **harness engineering** — agentic infrastructure is the next platform shift. Active portfolio: context engineering (ticket-rs), semantic code search (sgrep), natural-language data access (SQLGenie).

ticket-rs.io · Founder · 2026 → present

AI-native issue tracking. Git-backed, local-first, zero daemons. Treats the backlog as a graph; **PageRank-based prioritization** and **critical path analysis** over dependency edges to surface what to build next. First-class integrations with **Claude Code, Cursor, Codex, Copilot, Cline, Gemini, Windsurf, Zed, JetBrains AI, Goose** (MCP servers + Python SDK + GitHub / Linear sync). MIT-licensed, public alpha.

sgrep.sh · Founder · 2026 → present

Semantic search for codebases agents are writing faster than humans read them. Pure Rust, single static binary. 8M-parameter **Model2Vec** embeddings via **Candle** (HF's Rust ML framework); sub-10ms search, hybrid semantic + BM25 mode via ripgrep. MIT-licensed.

SQLGenie · Founder & CTO · 2024 → present

Natural-language query engine across **23+ SQL dialects**. Analyst workflows at **~200× lower per-query cost** than the manual baseline. Inbound from TechStars, multiple VCs, and two acquirers inside the first 90 days.

Hidden Layer Labs · Atom · CTO · 2020 – 2022

GPT-3-era personal AI assistant with data connectors across Spotify, Strava, calendar, and comms. Led a 5-engineer team; owned architecture, hiring, and product roadmap. Alongside: a lab of shipped micro-SaaS (*Theta Labs, Levitate, ML Academy, Prompt Fire*).

EXPERIENCE

AI Engineer · Strange Loop Labs

Feb 2025 → present · remote · US

"McKinsey — but instead of decks, we bring demos." Early engineer at a forward-deployed AI shop serving **Fortune 500** and **Big Four** clients (backed by Gaia Ventures, El Cap, AI Grant). Built by **Amazon Alexa** engineers. Own the technical architecture and delivery strategy for each engagement.

- End-to-end ownership — exec discovery through production. Full-stack **AWS** (Lambda · ECS · API Gateway · S3 · SQS): serverless and distributed systems, observability, SOC-2 Type 2 compliant deployments, runbook handover.
- Regulated enterprise document workflows for F500 / Big 4 clients — custom AI for the messiest unstructured data these organizations deal with, processing tens of thousands of forms at above-human accuracy.
- **Internal research: high-throughput agentic coding.** Applying classical ML techniques (ensemble methods, online learning, statistical quality control) to the code-generation loop — not just prompt engineering. Map-reduce-style ticket-driven development — structured variation on the **Ralph loop** with **ticket-rs.io** as the context graph. Agent harness ingests customer call transcripts and notes, decomposes them into dependency-ordered tickets, and dispatches coding agents that inherit the exact context the FDE heard. Result: features that previously required weeks of iteration now ship in days with full customer intent preserved.
- Technical lead on customer C-suite readouts; shaped hiring bar and technical interview loop as the team scaled.

Principal AI Scientist · Stealth AR/VR Startup

Oct 2024 → Apr 2025 · remote · US

Principal scientist on the company's sole product — a **manufacturing copilot** on **Meta Quest 3** and **Apple Vision Pro** delivering real-time assembly guidance to production-floor workers. Reported to CEO; two direct reports. Owned system architecture, technology selection, and go-to-market technical strategy.

- Shipped **YOLO-based real-time video pipelines** and vision-LLM systems that infer assembly workflows from training video and live camera feed — eliminated the per-customer setup cost that had been blocking deployments.
- Real-time multimodal inference (video · audio · text · spatial / LiDAR) on a multi-node GPU cluster; self-monitoring, canary rollout, automatic rollback.

Machine Learning Scientist · True Fit

Mar 2022 → Aug 2024 · remote · US

Owned the size-recommendation engine — True Fit's product — sizing millions of real 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 the largest cross-retailer e-commerce dataset on the web; served via **GCP Vertex**.

- Re-architected the recommender from legacy regression to a **3-class deep-learning ensemble**; shipped through dozens of A/B-tested release cycles.
- Integrated LLMs (GPT · Llama · Mistral · Gemini) and **BERT embeddings via Hugging Face Transformers** into product-attribute extraction, taxonomy classification, and reverse-image search; retired years of regex and manual curation.
- Rebuilt the A/B testing platform from parametric to non-parametric stats; eliminated the chronic false-positive rate that had been driving roadmap decisions — the company's product roadmap shifted as a result.
- Forward-deployed with retailer data teams; most shipped wins originated from those conversations, not from internal roadmap planning.

Data Analyst, Institutional Research · Westcliff University

Mar 2021 → Feb 2022 · Irvine, CA

- University outcome-prediction models (**>90% accuracy**); automated Python / SQL ETL across every major information system.

Technical Advisor · Topmate.io

2024 → present

- Architecture review, deployment strategy, and technical evaluation for AI/ML teams and early-stage founders.

WRITING & OPEN SOURCE

Author of **Vanishing Gradients** — technical blog on production AI, harness engineering, and the economics of coding agents. Thesis: *"The Harness Is the Product"* — 70% of agent failures are context errors, not model errors. Published in ***AI in Plain English*** and ***Level Up Coding***. **7 models on Hugging Face** — NeuralPipe (SLERP/TIES merges), NeuralHermes (GGUF quantization), Qwen2.5-Sci. Open-source: **gliner-as-a-service** (47 ★).

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

Agentic & AI — multi-agent orchestration · context engineering · RAG · evals · tool use · LLM fine-tuning (LoRA · SFT · DPO · RLHF) · multimodal (vision / audio / text / spatial) · real-time inference (YOLO · object detection) · on-device models · computer vision · NER · recommender systems · classical ML for code generation · ML/RL for coding agents

Frameworks — PyTorch · Hugging Face Transformers · LangGraph · MCP · vLLM · Triton · pgvector / FAISS · Claude Code · Cursor · Codex · Cline

Systems — AWS (Lambda · ECS · API Gateway · S3 · SQS) · GCP Vertex · Azure · Docker · Kubernetes · Terraform · stateless microservices · serverless · low-latency serving · GPU clusters · observability · MLOps · canary & feature flags

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