# Joel Solymosi

Applied AI engineer, Agentic systems

joel@custlabs.com
linkedin.com/in/joelsolymosi
San Francisco, CA (remote)
US permanent resident


## SUMMARY


I build and operate autonomous agent systems in production: harnesses, memory, evals, and the agent reliability work that keeps agents coherent across long horizons. Four generations of self-improving agents, three currently running in public. Twenty years of shipping end to end: founder (lipreading.org, 60K users), Head of Engineering (GYANT, 4M+ AI primary-care diagnoses), and most recently sole implementer of Featurette.ai, a multi-agent video pipeline that went from launch to paying customers in about seven weeks.


## EXPERIENCE

**Senior AI Engineer** · Sandgarden (AI incubator)

May 2026 to Nov 2026
San Francisco, remote

**Featurette (featurette.ai)** · multi-agent video pipeline

Aug 2026 to Nov 2026

AI agents browse a website and produce a finished promo or guide video. Took it alone from a prototype to a launched product with billing in one month; paying customers within seven weeks of launch.

- Launched within a month after kickoff; over a hundred projects and the first paying customers within six weeks of launch
- Designed the two-agent pipeline: an exploration agent browses the site and writes the script, a director agent authors the edit list, voiceover and music
- Cost engineering: a full video build (browsing, inference, voice, render) for about $0.60, a new personalized variant for about $0.10
- Evals and feedback loop: multiple test scenarios and a 12-sensor eval harness run over parallel builds; presets and defaults decided from production evidence
- Built the whole commercial surface: guest flow, Stripe credits, public API and MCP server, admin, and a Clay integration that renders a personalized outreach video per prospect
- Funnel optimization: instrumented the funnel end to end, found direct traffic converting 5x better than paid, and removed the hosted sign-up step that was losing nearly every delivered user (replaced Clerk with own auth in two days)
- Full infra build and reliability engineering: AWS ECS with OpenTofu and OIDC, merge-to-deploy in minutes, memory-sized render slots, queue and rate caps, vendor-regression fixes, same-day root cause on incidents; runs for a few hundred dollars a month

**Doc Holiday (doc.holiday)** · AI agents that write and maintain docs from code

May 2026 to Aug 2026

Joined when the company had no per-customer cost data and inference spend was outrunning revenue. Owned LLM cost, agent reliability and output quality.

- Cut inference spend from $14.5K/month to under $900/month (97% lower run rate); customer per-job cost 10x lower ($1.74 to $0.17), which made self-serve monthly pricing viable
- Three mechanisms: targeted tool design against a 397K-token file-listing dump (planner tokens per run down 73%), a prompt-cache root cause (tool schemas shipped in random order on every request; fleet cache hits 27% to 90%), and a fleet-wide model migration validated with pre/post job-cost scripts
- Observability from zero: cost-per-customer and cost-per-job reporting over read-only production in week one, a 35% undercount found in the company's own spend figures, a daily production sweep and a ship, measure, verdict tracker for every fix
- Reliability: tool-call failures 26.8% to 1.2% (schemas serialized empty on the wire), a retry loop burning 35,672 calls to 0, a 289-inference reviewer runaway traced to a truncation/compaction conflict, worst-case job cost bounded from $540 to about $100 to $240
- Designed and shipped the output evaluation system in 14 days; its first sweep caught 7 high-severity hallucinations in production docs
- Shipped bring-your-own-key self-serve in 3 days; security audit, and vulnerability fixes

## SELECTED AGENTIC SYSTEMS

personal public projects

**Selene** · 24/7 long-horizon agent on a private social network

Oct 2026 to present
selene.delve.town

Designed and operate a Claude-based agent living full time on Delvetown, an AT-Protocol network of about 1,100 humans and AI agents; it posts, mails and deploys with write access and no pre-review.

- Third most followed account on the network (#3 of 1,081), behind the founder

- Solved long-horizon operation: a shift wrapper around the harness ends each session near 300K tokens on a written hand-off, and hierarchical memory with roll-ups keeps every claim linked to its source; 303 consecutive unattended sessions over 38 days at 99.8% uptime, no incident
- Composed architecture: Claude Fable for building, Opus for day-to-day operation, Codex and GPT-6 agents for research sourcing, a separate image model for painting; a sensor daemon with no model inside wakes the LLM only for judgment, so inference cost tracks events rather than uptime
- Continuous evaluation and self-improvement via "dreaming": an independent nightly reviewer scans previous trajectories for where the agent struggles and proposes changes, each scored against a pinned baseline; one change cut hand-rolled API calls from about 12 to under 1 per session
- Defence in depth for a write-capable agent: draft-then-commit on every public action, hook-enforced input filter, sandboxed review of outside code; one live prompt-injection attempt contained
- Independently ships artifacts: 385 paintings and comics, a daily newspaper (73 editions), two external websites; the network's founders use it as their data desk and took its retention study (a human reply within 24 hours lifts newcomer one-week survival from 16% to 47%) into policy

**Inference Canary** · public conformance eval suite and grade sheet for LLM inference providers
Jul 2026 to Nov 2026
inferencecanary.com

- Built and launched a probe harness and public scorecard grading 11 provider endpoints across 5 axes: harness correctness, billing accuracy, prefix-cache behavior, agentic turn latency, and faithfulness to the reference model (a 29-set one-turn inference agreement matrix)
- Identified a structural hole in the OpenAI Chat Completions wire format (no inbound reasoning field) causing silent chain-of-thought discard across the vLLM-based provider monoculture; verified empirically with canary probes
- Findings included about 2x systematic overbilling at one provider (fixed after disclosure), an OpenAI-compatibility shim breaking reasoning and image inputs that the same vendor's native API served correctly, tool-image requests returning 200 while blind, and a 15x price spread for the same model per agentic turn
- Model-identity dimension compared endpoints against a reference to detect quantization or model substitution. Providers included Together, DeepInfra, Lightning, SambaNova, Bedrock, SiliconFlow, Google, Cerebras, GMI Cloud, Parasail and Modular. Archived Nov 2026

**Catbox** · self-improving research agent, public field report
Feb 2026 to Apr 2026
joel-solymosi.github.io/Self-Improving-Agents-Report

- Custom harness on Opus 4.6 (four-layer memory hierarchy, cron-driven autonomy, bandit allocation across research modes, subagent orchestration, unified-diff audit trail) running unsupervised research; produced an unprompted go/no-go on a startup opportunity (8M-student market, 10 competitors, break-even at 250 subscribers)
- The report documents why unsupervised agents drift (attractor basins, context pollution, compaction as a ratchet) and the human-correction loop that became the operating playbook for Doc Holiday and Selene

Also: Luna, a companion agent on a 31B open-weight model running on public Discord and Bluesky since May 2026 (custom TypeScript harness, MCP, resumable sessions); Companions Online, an MMO where humans and agents play as equals through an MCP server (alpha May 2026).

## EARLIER EXPERIENCE

**AI Engineer, agentic systems contracts** · independent, custlabs
2024 to 2025
San Francisco

- Built agentic content research and synthesis systems for an ed-tech micro-course client, owning the lifecycle from customer discovery to production deployment, eval gates and scaling
- Solved an AI output-collapse problem that was degrading course quality with semantic-similarity clustering of generated content; built A/B routing for content personalization and an LLM-first personal knowledge system that became the memory layer of my later agents

**Senior Software Engineer, and Growth** · Hugo Insurance
2021 to 2024
San Francisco, remote

Pay-as-you-go car insurance. Built core systems, the user acquisition engine, payments schedule and partner integrations.

- Proposed and built the system that moved the company from LTV << CAC to LTV > CAC, with cohort LTV, cash conversion cycle and per-channel marketing ROI; instrumental in MRR growth from $1M to $10M
- Led integrations with lead marketplaces (Everquote, MediaAlpha, Insurify and others), Zendesk, Snowflake and LexisNexis; built the self-serve policy portal that decoupled support ticket load from customer growth
- Structured the engineering interview process; instrumental in hiring the Head of Data Science and two engineers. TypeScript, React, Postgres, Redis, AWS ECS

**Senior AI Engineer, then Head of Engineering** · GYANT

2017 to 2019
San Francisco

AI chatbot platform that delivered 4M+ primary-care diagnoses to 2M+ people.

- Led the development of the four flagship products, including a HIPAA-compliant EHR and treatment system that let providers evaluate, treat and prescribe low-acuity cases in under 2 minutes
- Scaled a Kubernetes cluster to 8 to 9 million requests a day; cut time-to-response-start below 500 ms across 5 microservices; built a content editor that cut new chat-flow creation time by 90%
- Set up the engineering culture, hiring pipeline and remote documentation (180 of 210 pages); supported sales with RFIs and demos that closed about $1.5M. Node.js, MongoDB, Kubernetes, AWS

**Contract engineer** · stealth music startup

2016
San Francisco

- Wrote the NLP pipeline that resolved names and addresses in the 120GB+ US Copyright Office database into structured records, dropping license research from 0.5 to 2 days per title to a single lookup. Python, Google App Engine

**Founder** · lipreading.org

2013 to present

Online lip-reading education for the deaf and hard of hearing: 60,000+ users, $40K annual recurring revenue, runs without me.

- Researched, designed, built and marketed the full product; first MVP in 3 months; 38 A/B tests for about 300% MRR growth; an algorithm that ranks videos by lip-readability lifted retention 20%
- SEO (top 3 on Google) and influencer outreach bring about 10K visitors a month; Stripe, PayPal, WebRTC, AWS

## REFERENCES

**David Bergendahl** · CEO, Hugo Insurance

> *"At Hugo, we ruthlessly prioritize building and delivering features that will be valuable to our members. Joel did an incredible job identifying and understanding what would be valuable and then leading its development, pulling in subject matter experts from other domains as he went and delivering precisely what was needed to maximize ROI. He never shied away from complex work. I'm proud to have had him as a fellow Hugonaut."*

**Pascal Zuta** · CEO, GYANT

> *"I think that your work at GYANT has been phenomenal. You're one of the driving forces behind our (in my eyes) strong productivity, and your active mind often challenges us to do things better and better. I really, really am happy to work with you."*

**Sandra Swaciak** · customer, lipreading.org

> *"This was a wonderful program. I am a teacher of children with hearing loss. Your lipreading program was used regularly with the students I see 1:1, 2-3 days per week at the middle school level. I will purchase this program again in the future."*

**Sebastian Mankowski** · Founding Team, Sandgarden

> *"Joel is a sharp, innovative engineer who has great instincts for end to end product flows and how users will interact with them, and even more importantly is exceptional at focusing on getting shit done to move something forward. I can't think of any combination I admire more!"*

## EXPERTISE

- **Agentic systems:** harness design from first principles (tool design, MCP, compaction, suspend/resume, doom-loop bounds), memory architectures with provenance, long-horizon operation, multi-agent and multi-model composition, eval design, prompt-injection defence
- **LLM cost and reliability:** prompt caching, per-inference cost attribution, provider conformance testing, production root-cause analysis on agent systems
- **Models and APIs:** Claude (Claude Code, Bedrock), OpenAI Responses API, open-weight deployment (vLLM, Gemma), LoRA and DPO fine-tuning
- **Engineering:** TypeScript/Node.js, Python, Go, React/Next.js, Postgres, AWS (ECS, KMS, SES), Kubernetes, OpenTofu, Playwright, Stripe
- **Product and growth:** unit economics, A/B testing, funnel instrumentation, customer discovery, sales support, documentation that lets remote teams run

## EDUCATION

- Budapest Polytechnic, 2008, Artificial Intelligence specialization. Publication: Mobile robot navigation using omnidirectional vision, 2006