

Joey Hersh

Engineer III - AI candidate · Agentic Systems Engineer

San Antonio, Texas | me@joeyhersh.com | 713-505-3289 | <https://github.com/joey12725>

SUMMARY

Agentic systems engineer with regulated banking delivery experience, applying for Frost Bank Engineer III - AI (R260062). I design and operate agent-loop architectures with tool orchestration, planner/verifier separation, memory and context handling, caching, bounded concurrency, and fallback strategies. Shipping credit-risk data systems inside a large bank since 2023 with validation, lineage, change control, and mentoring. Python, FastAPI, LangChain, LangGraph, vector databases (pgvector, ChromaDB, Pinecone), AWS, event-driven and API orchestration, Weights & Biases and Weave evaluation.

KEY NUMBERS

- ≈26M records a day, through production data pipelines at USAA
- ¼ trillion+ tokens processed, input plus output, across every model family since GPT-3 (modeled estimate)
- 40,000 campaign leads, at about \$0.29 per lead for Fit & Flourish
- \$100K+ funding and prizes, for ReCap Budget, \$46K cash plus about \$60K in-kind

EXPERIENCE

Data Engineer III, USAA

June 2024 – present | San Antonio, Texas

- Delivered complex data engineering features supporting credit-risk analytics, regulatory compliance, and a major FICO model migration using Snowflake, dbt, Python, SQL, GitLab CI/CD, and Control-M.
- Backfilled and corrected 19 years of historical consumer-lending data (2006-2025) using targeted validation and correction rather than rebuilds; maintained zero downtime through legacy retirement and model transitions.
- Delivered a FICO 10T data asset supporting CECL by the September 2025 regulatory deadline and retired a legacy consumer-lending pipeline with uninterrupted reporting.
- Built reusable data-validation patterns and troubleshooting documentation; added scheduler dependencies that reduced batch-failure and manual-intervention risk.
- Mentored and onboarded engineers across dbt, Snowflake, scheduling, testing, and access; authored dozens of reusable SKILL.md guides; taught AI-assisted development practices across the team.
- Member of the Data Science and Engineering Guild contributing to decisions and education on generative-AI adoption across the bank; manager-documented progression toward independent ownership of complex regulated work.

Founder and CEO, Guadalupe Automation

January 2026 – present | San Antonio, Texas

- Founded and lead an AI automation company delivering agentic systems for small and mid-sized businesses, from discovery and workflow modeling through architecture, implementation, deployment, and handoff.
- Architected a retail accounting-operations platform separating deterministic reconciliation logic (Python) from LLM-based document extraction with human exception review, audit state, and requirements traceability.
- Designed a public-record enrichment and lead-qualification system with modular data providers, configurable scoring, provenance, confidence, and explainable call rationale.
- Operate isolated per-client agent deployments (VPS plus preconfigured Docker stacks) and separate containerized environments for private users, including workspaces, browsers, credentials, and schedules.

Co-founder, marketing and automation lead, Fit & Flourish

2024 – 2025 | San Antonio, Texas

- Co-founded a consumer fitness company; led marketing automation, analytics, e-commerce, referral, and tracking systems.
- Built an autonomous ad-creative testing system (Meta Ads API, n8n, image generation, multimodal review, human approval) that generated 40,000 waitlist leads at approximately \$0.29 cost per lead.
- Hired and led a team of six; ran API-based analytics and A/B testing.

CEO and Machine Learning Engineer, ReCap Budget

December 2021 – September 2023 | San Antonio, Texas

- Founded and led a seven-developer team building a serverless personal-finance application on more than 25 AWS services; raised \$46,000 in seed capital and approximately \$60,000 in in-kind prizes.
- Built an event-driven receipt-processing pipeline (S3, Lambda, Textract) and a fine-tuned GPT-3 line-item classifier covering 100+ categories with over 85% reported accuracy.
- Implemented structured-output frameworks for LLM responses, FastAPI services, model evaluation, and bias-mitigation work on a fine-tuned advice-generation model.

Data Engineering Intern, USAA

June 2023 – August 2023 | San Antonio, Texas

- Delivered a credit-card underwriting structured-view feature from requirements to production release; built Python unit-test and data-validation scripts; completed architecture review and production control gate.
- Built and presented an internal vector-database-backed (ChromaDB, Falcon-40B) retrieval-augmented generation proof of concept to hundreds of employees; helped five colleagues deploy the code in an internal environment.

FROST BANK ENGINEER III - AI (R260062) ALIGNMENT

- **Design intelligent systems and workflows that use agentic behavior and LLM decision-making:** Joy, my store-close reconciliation platform, and SecondHome each decompose a real workflow into deterministic code, model judgment, durable state, and human review. I've done it for my own operations, for a retail accounting client, and for a real-estate lead pipeline.
- **Drive design sessions for agent loop architectures, tool orchestration, and task planning:** Joy's overnight pipeline has explicit planner, worker, verifier, and recovery stages. Cyrus-v2 is a custom orchestration platform I built to run coding agents against real backlogs with budgets, schedules, and a compliance gate. Guadalupe engagements start with architecture sessions I run.
- **Lead technical efforts on AI reliability, bias mitigation, and drift detection:** Reliability is my strongest suit: retries, leases, fail-closed defaults, evidence-bound completion. I did bias-mitigation work on ReCap's fine-tuned advice model. I evaluate agent behavior with Weave and W&B. I won't claim a formal production drift program because I haven't run one; I have run the regression loops that would feed it.
- **Evaluate agent behavior for fairness, reliability, explainability, and safety:** SecondHome puts a written rationale and confidence on every score. GBrain stores cited knowledge with temporal boundaries. Morning operations report what was learned, not just that they ran.
- **Optimize agentic systems with caching, throttling, and fallback strategies:** Prompt caching, model-by-task routing, bounded concurrency, and fallback chains are how Joy stays under budget and keeps working when a source fails. One environment logged over a thousand scheduled runs in eighteen hours.
- **Event-driven design, API orchestration, memory and context handling:** ReCap's receipt pipeline was S3 to Lambda to Textract to a fine-tuned classifier in 2022. GBrain is a pgvector-backed memory layer with graph edges and source chunks. Cortex triages audio locally before it ever calls a cloud model.
- **Python, FastAPI, LangChain, vector databases, and agent loop architectures:** Python daily for years. FastAPI at ReCap, in Cortex, in SecondHome's API, and in current private services. LangChain in the offline RAG prototype and a pile of experiments. pgvector in production for GBrain; ChromaDB and Pinecone before it. LangGraph used. I only list frameworks I've actually run.
- **Mentor engineers and collaborate with product and design on outcomes and metrics:** Manager-documented mentoring, onboarding, and reusable guides at USAA. Seven-developer team at ReCap. Client discovery and outcome definition at Guadalupe. An internal GenAI talk to hundreds of employees as an intern.
- **Enterprise-grade AI/ML experience and regulated banking context:** Building with language models since 2019. Inside a bank's credit-risk data team since 2023, with validation, lineage, controlled change, and regulatory deadlines. I already work inside a bank. The open question is whether I can bring agent architecture in without breaking the controls. That's the part I like.

SELECTED SYSTEMS

Joy (2025 – present)

Joy is my agent. Hermes runs its loop. GBrain holds durable, cited memory on pgvector. A self-hosted Plane instance is where Joy and I hand work back and forth, with evidence attached to every task. A custom model router, fine-tuned on a 4-

bit Qwen3 8B, decides which model each job deserves. A staged overnight pipeline plans, dispatches, verifies, and refuses to call anything done without evidence. Role: Architect, operator, and product owner. Much of the code is agent-authored under my direction and review.

Regulated credit-risk delivery (June 2023 – present)

Historical backfills across nineteen years of data. A legacy pipeline retired with reporting uninterrupted. FICO 10T and CECL assets delivered on the regulatory deadline. Dozens of SKILL.md guides the team runs on. A seat in the guild that decides how the bank uses generative AI. Role: Data Engineer III on a credit-risk data team, after a 2023 internship on the same team. Technical ownership, mentoring, and cross-team coordination.

Store-close reconciliation (June – September 2026)

A retail accounting-operations platform. Fifteen deterministic parsers read bank statements, point-of-sale exports, and commission and discount workbooks. A model lane reads only the unstructured evidence, against a closed set of schemas, and everything it proposes goes to a human queue. Reconciliation is Python. Exports require a role-gated approval. The June close reconciled 295 of 315 checks with zero open amount differences. It also taught me what "done" means to a client. Role: Architect and implementation lead at Guadalupe Automation, with a partner running the client relationship. Built with coding agents I directed and reviewed.

Cyrus-v2 (March 2026)

Cyrus-v2 forks an open-source project tracker and turns work items into dispatchable agent sessions. A relay normalizes Claude Code, Codex, Gemini CLI, and OpenCode behind one event stream with per-model token and cost accounting. A metering engine enforces compute-pool budgets. A scheduler with cooldowns and circuit breakers decides what runs. A lifecycle layer audits each work item for compliance before it can be called done. An autonomous PM agent grooms the backlog between dispatches. Role: Sole human author. Designed the relay, orchestrator, metering, scheduling, and PM agent on top of a forked project tracker, and built them with coding agents I directed and reviewed.

SecondHome (July 2026)

Distressed-property lead discovery across twenty county sources. Records are normalized with their source kept on every field, enriched through swappable providers, researched by bounded agent skills, and scored by a deterministic, versioned formula. Contact channels need two independent signals before they count. Every lead ships with a written call rationale. The product didn't find its market. The architecture is still the one I'd use. Role: Product lead and sole author. 223 commits in a twelve-day sprint, working with a real-estate operator as the intended user.

EDUCATION

B.S. in Computer Science, Trinity University, San Antonio, Texas, 2024

SKILLS

What I reach for first: Python, SQL, Hermes, Postgres + pgvector, DuckDB, AWS Lambda, Docker, FastAPI, Claude Code, Codex, Cursor, OpenCode, Vercel

Used in systems that had to work: Snowflake and dbt, GitLab CI/CD, Control-M and DataStage, LangChain, LangGraph, ChromaDB, Pinecone, Redis, n8n and Make, AWS Bedrock, Weights & Biases and Weave, PyTorch and TensorFlow, ComfyUI, OpenClaw

Built or materially extended: Joy, Hermes skills, and morning operations, GBrain and the Living Atlas, Cyrus and Cyrus-v2, TinyVDB, Splice, Cortex, VibeTest, Spec-driven data-migration agents, Reusable SKILL.md guides