

Ian Greenberg

AI Engineer — LLM Agents, RAG & Inference

Dallas, TX · +1 214-603-7013 · ian@iangreenberg.com · iangreenberg.com · github.com/iangogentic · linkedin.com/in/ian-g-390487b6

SUMMARY

Applied AI engineer who ships LLM systems into production for non-technical operators — retrieval augmented applications, multi-agent workflows over live company systems, and approval-gated autonomy. Sole engineer on a company-owned platform running in production behind authentication, and author of an open patch to the vLLM inference server.

EXPERIENCE

Digital Marketing Manager (Sole Engineer, Internal Platform) — SIG ORG / SIG University Jul 2024 – Apr 2026

- Sole engineer on a company-owned email operations and analytics platform — **194 of 195 commits** across the repository's lifetime — running in production on Vercel behind Clerk authentication.
- Integrated Pardot, Salesforce, Cvent, Stripo and Smartsheet behind one interface, covering 200+ Pardot campaigns documented in the platform's configuration, and ran nine scheduled cron routes in production.
- Shipped LLM features into that platform: OpenAI-powered agent and chatbot functions, and Gemini-based retrieval over campaign and content data.

Director of AI & Automation — Z-School May 2026 – Jul 2026

- Built the company's agent platform on custom MCP servers over Zoho CRM, Calendly, WorkDrive and MS Graph, deployed as scheduled production services with consequential actions gated on explicit human approval.
- Designed a multi-agent outbound system — sender, supervisor, top-up and reply-watcher services with suppression rules, sender health monitoring and a kill switch — running a reactivation program to ~320 lapsed customers across ~500 tracked sends and recovering \$75K in attributed revenue.

Co-Founder & AI Engineer — Gogentic (co-owned) Feb 2025 – Present

- Build agent harnesses, retrieval systems and local-inference tooling, including the agent systems deployed at Z-School above.
- Built waterfall contact enrichment via the Apollo API into Gogentic's automated outbound email system.
- Develop Cortex, automated course-creation software licensed to ~20 universities through partner networks.

Co-Founder — GoGogiri LLC May 2021 – Jun 2024

- Co-founded a Japanese-Texan onigiri concept and grew it from one rented kitchen slot to three concurrent locations and 30+ employees; press coverage in CultureMap Dallas, the Dallas Observer and CW33.

SELECTED PROJECTS

GPU Cats (C++, macOS) — desktop application that renders coding agents as animated characters, orchestrating work across isolated git worktrees with local voice and persistent memory. 1,300+ commits, in private beta. gpucats.iangreenberg.com

localnpc-engine (Rust, MIT) — engine for running local models on real tasks: tool calling, replayable traces, evaluation harnesses, and a completion gate requiring executable evidence before a run can finish. github.com/iangogentic/localnpc-engine

Ian Brain (MCP) — private gateway over machines, memory and files across ten capability tiers, with an out-of-band approval path outside the model's tool surface. Hybrid retrieval over the memory store using BGE-M3 embeddings, reciprocal rank fusion and a local cross-encoder reranker. ianbrain.iangreenberg.com

OPEN SOURCE & RESEARCH

- **vLLM** — author of open PR #46218: dtype fix in the AWQ/GPTQ path for MLA attention.
- **Inference benchmark** — ~257 tok/s single-stream decode on Qwen3.6-35B-A3B (3B active, AWQ 4-bit) on one A100-80GB, ~1.9× over the non-MTP baseline via MTP self-speculative decoding; internal reproduction notes.
- **Affiliated Researcher, Center of Creative Computation, SMU** — M3 cluster and DGX SuperPOD access through 2027; independent work on serving optimization and mixture-of-experts pruning.

SKILLS

LLM & agents: RAG (retrieval augmented generation), embeddings and vector search, prompt engineering, MCP tool calling, structured output, agent loops, multi agent orchestration, context engineering, persistent memory, human in the loop approval gates, evals and LLM as judge, guardrails, OpenAI, Anthropic Claude and Gemini APIs

Inference & retrieval: vLLM, SGLang, llama.cpp, Ollama · quantization (AWQ, GPTQ, GGUF), speculative decoding, KV cache quantization, mixture-of-experts pruning, throughput and latency benchmarking, NVIDIA DGX and A100 · BGE-M3, FAISS, hybrid retrieval, reciprocal rank fusion, cross encoder reranking

Engineering: Python, TypeScript, Rust, C++, SQL · FastAPI, React, REST and JSON APIs, PostgreSQL, SQLite, Docker, Linux, Vercel, Cloudflare, git, CI/CD, observability, threat modeling for agentic systems (prompt injection, tool permission scoping, gateway ACLs)

EDUCATION

Bachelor of Science, Human Sciences — Texas Tech University, May 2023