David Montgomery

(303) 335-7057 | DMontg@gmail.com | Centennial, CO

AI Innovation & Systems Engineer

GitHub | Portfolio | LinkedIn | Medium — 758 Stars • 158 Forks • 21K+ Downloads

TECHNICAL PROFILE

I find problems that AI probably can solve, figure out whether it actually can, and build the thing to find out. That has led me into computer vision, NLP, RAG systems, model fine-tuning, multi-LLM orchestration, edge inference, and security—not because I planned a broad curriculum, but because interesting problems kept showing up. I have shipped production systems across most of those domains, and I have a senior management background that means I build for real adoption, not just demos.

- Breadth of application: vision models trained for nighttime security camera footage; 6-model AI deliberation councils for clinical brain health analysis; audio-to-spectrogram species classification running on a Raspberry Pi; a full RAG platform with integrated model training studios, knowledge graph retrieval, and enterprise alerting—all built because I got curious and started building
- Full model lifecycle, end-to-end: from domain-specific dataset curation and LoRA fine-tuning through training visualization, evaluation frameworks, and automated regression analysis that tells you exactly which config change moved which metric
- Former senior manager of a 12-person team—understand how AI tools need to be designed for the people who will actually use them, not just the engineers who build them

CORE TECHNICAL SKILLS

Computer Vision & Edge AI

- Object detection: YOLO architectures (9T, 9C, 9C-320), domain-specific fine-tuning (security camera footage, avian species), OpenVINO, CoreML, TensorFlow Lite
- Color space and preprocessing optimization for IR/nighttime footage; multi-stage classification pipelines combining visual, temporal, and geographical context signals
- NVR integration: Scrypted, Frigate, RTSP/RTMP stream processing, plugin development; edge inference on Intel NPU/iGPU, Coral TPU, Apple Neural Engine

Model Training & Evaluation

- LoRA fine-tuning (rerankers and agentic LLMs) with MLX on Apple Silicon; cross-encoder training including triplet mining, pairwise conversion, and balanced negative sampling
- Custom training visualization: real-time parameter space trajectory rendering, live loss/gradient telemetry via SSE
- Evaluation frameworks with 20+ ranking metrics (MRR, Recall@K, NDCG, precision, latency percentiles); automated regression analysis that identifies which specific config change moved which metric and why

RAG, LLMs & Search

- Hybrid retrieval: vector (pgvector/Qdrant), sparse (BM25/FTS), and graph (Neo4j) with RRF fusion; knowledge graph construction, entity extraction, community detection
- Multi-LLM orchestration: parallel deliberation councils, cross-model comparison and consensus consolidation, Model Context Protocol (MCP) server development
- LLM integration: OpenAI API (Responses API, Eval SDK, Agents SDK), Anthropic, Google; embedding models: OpenAI, Cohere, Voyage, BGE, sentence-transformers

Languages & Frameworks

- Languages: (Py)Torch, Python, TypeScript, JavaScript, SQL, Bash, Go (exporters/glue)
- Backend: OpenAPI, FastAPI, Node.js/Express, Flask, asyncio, multiprocessing
- Frontend: React, Vite, TailwindCSS, Dockview, Electron, Streamlit (for dev)
- Data: Pandas, Pydantic models, ETL pipelines, tree-sitter AST parsing

Infrastructure & DevOps

- Cloud: AWS (S3, EC2, Lambda), Netlify, serverless architectures
- Containerization: Docker, Docker Compose multi-profile deployments, Proxmox (~25 VMs in home lab)
- Monitoring: Prometheus, Grafana, custom hand-built Prom exporters for: Intel NPU, iGPU, MacMon; Telegraf; extremely proficient with Linux and command line

- Databases: PostgreSQL (pgvector, ts_rank), Qdrant, Neo4j, Redis, SQLite, MongoDB

Security & Compliance

- HIPAA: PHI handling, HMAC webhook verification, compliant logging, short-TTL tokens, zero PHI in logs
- Authentication: API key management, JWT, OAuth2, TBAC/RBAC, hierarchical multi-tenant structures
- Penetration testing: Nmap, Metasploit, Burp Suite, Kali suite, Pentest MCP (creator)

PROFESSIONAL EXPERIENCE

Neuro-Luminance Center for Brain Health ***May 2025 – Present***

AI / RAG Engineer

Architect, build, and operate production RAG, analysis, and patient communication platforms for clinicians and researchers interpreting qEEG brain health data.

qEEG Council — Multi-LLM Deliberation Platform

- Built 6-stage deliberation workflow with parallel analysis by Qwen3, GLM-4.7, DeepSeek-3.2, GPT-5, Opus 4.5, and Gemini 3 Pro; vision-first approach rendering PDF pages as images for multimodal understanding of charts, graphs, and brain maps
- Cross-model comparison stage identifies agreements and flags discrepancies for human review with confidence scoring; consensus consolidation reduces hallucinations in clinical recommendations through structured verification
- Designed end-to-end RAG pipeline using PostgreSQL/pgvector for RRF-based hybrid and multi-modal search over clinical notes, qEEG interpretations, and neuroscience literature (text + code + image + 3D file inputs)
- Ingest and normalize qEEG reports, PDFs, and EDF (European Data Format) files into structured representations with per-region metrics and event-related potentials (ERP); full deterministic data-trap structure with QC gates in every stage

Patient Explainer Video Pipeline (local-explainer-video)

- Built local, offline-capable alternative to NotebookLM for converting qEEG brain scan reports into patient-friendly explainer videos with scene-by-scene editability
- Multi-stage pipeline: LLM director agent breaks clinical text into 5-15 scenes with narration and image prompts, generates voiceover (Kokoro local TTS / ElevenLabs / OpenAI), and assembles final MP4 with MoviePy/ffmpeg
- QC verification gate: judge model (Claude Opus) validates narration accuracy against qEEG Council ground truth artifacts; Gemini vision detects misspellings and data errors in rendered slide images
- Automated image editing via Qwen Image Edit for slide text corrections without full regeneration; publishes verified MP4s to clinician portal sync folder with DB-tracked file uploads

Thrylen Platform & Medical Document Handling

- Implemented AWS S3 integration for secure document storage of patient reports, consent forms, and longitudinal qEEG data
- Metadata tagging for downstream RAG: pseudonymized patient IDs, timestamping, modality labels
- Safety-first: limited scope prompts, explicit documentation of system capabilities and limitations

RagWeld — Production RAG & AI Training Platform ***2024 – Present***

Founder & Lead Developer

Full-stack RAG platform combining vector, sparse, and graph search with integrated model training studios, evaluation infrastructure, and enterprise-grade observability. Designed as both a production system and an educational knowledge suite for learning RAG concepts. Pydantic-first architecture with 1,160+ tunable fields auto-generating TypeScript types and enforcing validation at all boundaries.

Tri-Brid Retrieval & Knowledge Graph

- Three parallel retrieval engines fused via configurable RRF or weighted scoring: pgvector dense search (HNSW), PostgreSQL full-text BM25 sparse search, and Neo4j graph search with entity extraction, relationship mapping (calls, imports, inherits, references), and community detection (Louvain, Label Propagation)
- Knowledge graph construction: automatic entity extraction (functions, classes, modules, variables), graph inspection UI with entity search, neighbor subgraphs, and community browsing; read-only Cypher query endpoint for debugging

- MCP integration: embedded Streamable HTTP server with search, answer, and list_corpora tools—Claude Desktop / IDE ready, stateless HTTP mode, configurable auth

Chat with Persistent Recall

- Conversational RAG chat with persistent vectorized memory: every conversation auto-indexed into pgvector + FTS with configurable chunking strategies (sentence, turn, paragraph, fixed)
- Intelligent retrieval gate: pattern-based classification of user messages into recall-triggering (past references, shared context) vs standalone questions, with configurable intensity levels and fusion overrides per recall plan
- Per-request tracing with debug footer showing confidence, active retrieval legs, fusion method, result counts per leg, and run ID; Loki log integration with SSE live streaming in the Chat UI

Model Training Studios

- Dual LoRA training studios (React/Dockview + FastAPI/MLX backend): Qwen3 cross-encoder reranker fine-tuning (0.6B) and Qwen3 agentic retrieval LLM training, each with 40+ configurable hyperparameters and real-time visual monitoring
- Visual gradient descent: 2D projection of LoRA parameter space trajectory via Gram-Schmidt orthogonalized basis vectors, step-level SSE telemetry (loss, gradients, projections, learning rate); NeuralVisualizer and GradientDescentViz React components for interactive exploration
- Triplet-to-pairwise training pipeline with configurable negative ratio capping (5:1), deterministic train/dev splitting, auto-promotion on metric improvement, and auto-rollback mechanisms

Evaluation & Analysis Engine

- Comprehensive eval framework: 20+ ranking metrics (MRR, Recall@5/10/20, Precision@5, NDCG@10, Top-1/K accuracy, latency p50/p95) across 1,160+ tunable configuration parameters in 20 sections
- Eval Drilldown: side-by-side run comparison showing exactly which of 500+ parameters changed, with before/after values and per-question metric deltas
- AI analysis pipeline: computes metric deltas, identifies per-question regressions and improvements (up to 25 examples), routes to LLM for root cause hypotheses, suggested config knobs, validation steps, and confounding variable warnings

Enterprise Observability & Alerting

- Prometheus metrics + Grafana dashboards + Loki log aggregation; per-request cost breakdowns (generation + embeddings + reranking) with daily/monthly projections
- Custom webhook alerts to Slack and Discord with configurable severity (critical/warning/info), resolved notification support, and configurable timeouts—production-grade canary monitoring for MRR drops, quality regressions, or any user-defined threshold

Educational Knowledge Suite

- 250-term glossary with contextual tooltips throughout the UI (7,700+ line glossary.json with definitions, related terms, external links, and expert/warn badges)—the application itself is designed as a comprehensive learning tool for RAG concepts and architecture

Computer Vision & Edge AI Projects ***2023 – Present***

Independent Developer & Collaborator

Security Camera Vision Model Optimization (with Koush / Scrypted)

- Collaborated with Scrypted NVR creator (Koush) to optimize YOLO object detection models specifically for security camera footage—an underserved domain where standard training datasets (daytime, high-resolution, RGB) fail on real-world conditions
- Achieved YOLO9T-320 outperforming YOLO9C-320 by ~30% on nighttime B&W human/animal detection accuracy in both OpenVINO and CoreML—a counterintuitive result demonstrating how domain-specific optimization can outweigh model size
- Reduced false positives from environmental triggers (wind, swaying branches, shadows) by ~42% through targeted training data curation and model tuning
- Converted input pipeline from RGB to YCbCr color space for improved performance on IR and grainy nighttime footage where chrominance data is noise rather than signal
- Long-distance person vs animal detection optimization for challenging low-visibility security conditions; built comprehensive benchmark suite testing inference across OpenVINO, TFLite, and CoreML backends

Bird Species Classification & Environmental Monitoring

- Fine-tuned custom YOLO models for avian species identification using CalTech and iNaturalist datasets; discovered unique challenges of avian CV (seasonal molts, near-identical species, hybrid edge cases)
- Implemented multi-stage classification pipeline combining temporal, geographical, and visual features—significantly improving accuracy over standard object detection by accounting for seasonality, location sensitivity, and complex lighting conditions
- Built BirdNET Scrypted plugin integrating TensorFlow Lite audio-to-spectrogram classification with virtual camera devices and configurable RTSP/USB audio input
- Created MCP server (mcp-server-birdstats) integrating eBird and BirdWeather APIs with dual transport (stdio + Streamable HTTP), token-optimized tool design, and structured error handling—enables natural language ornithological data analysis across ~1M annual detection points

FaxBot — Healthcare Integration Middleware Jan 2024 – Present

Founder & Lead Developer

First and only open-source, self-hostable fax platform with production-ready features and AI assistant integration.

AI & MCP Integration

- Three MCP servers (stdio, HTTP, SSE) enabling natural language fax operations with Claude Code and other AI assistants; identical SDKs for Node.js and Python published to npm and pip
- Traits-first architecture with feature gating—UI and API dynamically adapt to active provider capabilities

Technical Architecture

- Backend: FastAPI with asyncio for concurrent webhook handling, Redis for state, SQLite for metadata; multi-backend adapter pattern: cloud providers (Phaxio, Sinch, Documo, SignalWire) and self-hosted (FreeSWITCH, SIP/Asterisk)
- HIPAA-aligned: HMAC webhook signature verification, short-TTL API tokens, zero PHI in logs, multi-key API authentication with scoped permissions
- Reduced webhook processing latency by 40% through async/await optimization; automatic provider failover with circuit breaker pattern for 99.9% uptime; 85% code coverage

Jobot LLC Dec 2021 – Sep 2024

Senior Manager & RAG Systems Developer

Hybrid role: led 12-person recruiting team while building enterprise RAG system for AI-assisted recruiting.

RAG Platform for High-Volume Recruiting

- Designed and implemented enterprise RAG pipeline processing 2M+ candidate profiles and 10K+ client company records
- Led migration from ChromaDB prototype to Qdrant—improved query latency, enabled hybrid search with flexible filtering
- Built semantic job-candidate representations enabling natural language search, context-aware recommendations, talent rediscovery; OpenAI embeddings integrated with custom CRM via REST and GraphQL APIs

Fraud & Security Detection

- Architected pattern-based security scanning for phishing, fraudulent job postings, suspicious applications; heuristics: unusual domain/email patterns, reused scam language, abnormal geography/pay/benefit claims

Leadership

- Managed 12-person recruiting team: hiring, coaching, performance management, process improvement; recognized as top performer among 800 recruiters—systems built were actually usable by front-line staff

Earlier Experience:

Experis Finance 2018–2021 (Executive Recruiter)

Beacon Hill 2016–2018 (Senior Consultant)

Robert Half 2014–2016 (Staffing Manager)

Donor Development Strategies 2007–2014 (Co-Founder)

—Progressive leadership in technical recruiting; developed matching algorithms that informed later RAG embedding design

ADDITIONAL PROJECTS & OPEN SOURCE

- MCP 3D Printer Server: MCP server bridging AI assistants to 7 printer management systems (OctoPrint, Klipper/Moonraker, Duet, Repetier, Bambu Labs via MQTT/FTPS, Prusa Connect, Creality); STL manipulation tools (scale, rotate, translate, section-specific modifications, SVG visualization), Blender bridge, and end-to-end STL→slice→print workflow; dual transports (stdio + Streamable HTTP), Docker support, behavior tests; published on npm
- TTT-SSM Eval: Research infrastructure for safe Test-Time Training in State Space Models—fast weights update per-session during inference while core model weights remain frozen; defense-in-depth safety stack (entropy gate, canary loss probes, rollback with transaction semantics, Safety-Projected Fast Weights constraining gradients into safe subspaces); session forking for counterfactual comparison; architecture inspired by Complementary Learning Systems (fast hippocampal learning, slow neocortical consolidation)
- Vivified Bootstrap: Polyglot plugin kernel bridging Python and Node.js with JSON-RPC IPC, trait-based capability management, and type-safe plugin protocols; includes lightweight AI studio
- OpenPilot / Comma.ai: Active exploration of autonomous driving systems—object detection, language model integration, and security implications of locally-running inference on SnapDragon 845
- Pentest MCP: Created Model Context Protocol server for penetration testing workflows with Claude Code
- SecurityLens: Security tooling and vulnerability analysis platform
- Intel NPU/iGPU Prometheus Exporters: Custom-built monitoring exporters for Intel NPU and iGPU telemetry, macOS monitoring via macmon-prometheus-exporter; Grafana dashboard integration
- Home Lab: Proxmox cluster (~25 VMs), dual Scrypted NVR instances, Frigate, Pi-Hole, pfSense, Home Assistant, n8n automation—production-grade infrastructure for AI experimentation

EDUCATION & CERTIFICATIONS

- University of Denver CE – AI for Cybersecurity (Oct 2024 – Jun 2025)
- M.S. Organizational Development – Colorado State University
- B.A. Political Science – Metropolitan State University of Denver
- CompTIA: Security+, Network+, Linux+, A+ (in progress)