David Montgomery

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

Full-Stack AI / RAG Platform Engineer

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

TECHNICAL PROFILE

Full-stack engineer specializing in retrieval-augmented generation (RAG), GraphRAG, AI enablement platforms, LLM Training and Evaluation, and enterprise integrations. Blend of senior leadership background and hands-on engineering across RAG systems, observability, security tooling, and multi-agent LLM workflows.

- Design, build, maintain, break, fix, and love, production RAG systems that non-technical users rely on daily: recruiters, clinicians, operations staff. I strongly believe that breaking Retrieval at scale is the most valuable experience one can have in this domain.
- Deep experience turning messy real-world data (faded schematic pdfs, large-format blueprints, medical scans, resumes, job descriptions, medical reports, codebases) into usable vectors and knowledge graphs
- Strong alignment with SpectrumGPT-style work: Python/Node.js services, TypeScript, REST & jQuery APIs, vector DB (Qdrant, pgvector, Chroma), graph DB (Neo4j, GraphRAG, LightRAG)
    - OpenAI Apps SDK / Agents SDK (built one of the first 100 published plugins for gpt-3.5), AWS, microservices
- Former senior manager of a 12-person team—design AI tooling with adoption, workflows, and business logic, not just model benchmarks. I understand the what/how/why/when/who of separate use cases for AI tools across an organization; I understand how a single endpoint or tool call to the ML team, means something completely different, to every team across the organization
-

CORE TECHNICAL SKILLS

RAG, LLMs & Search

- RAG architectures: RRF-based hybrid lexical + vector search, multi-stage retrieval, self-learning cross-encoder reranking (e.g., using SBERT/SentenceTransformers), and chunking and tokenization strategies optimized by data type (e.g., using stemmers, whitespace tokenizers, or WordPiece models).
    - Experience and understanding to know when a 12-stage pipeline with KG and re-rankers on both sides of the gates is a valid design; and to know when grep/semgrep/ripgrep, or vanilla BM25 are the more valid choice
- Vector and graph databases: Qdrant, PostgreSQL/pgvector, Neo4j, ChromaDB, Redis
- Hybrid search: Triple-weighted RRF fusion across BM25, PostgreSQL ts_rank, and dense vectors (pgvector + Qdrant); entity-aware chunking, per-corpus retrieval strategies, multi-query and dual rerank pipelines
- Graph RAG: Microsoft GraphRAG, LightRAG, Neo4j—knowledge graph construction, entity extraction, community detection, and graph-augmented retrieval pipelines
- LLM integration: OpenAI API (Responses API w/ Lark + CFG), chat_completion, embeddings, Eval SDK, Apps/Agent SDKs), multi-LLM pipelines for cross-checking, Model Context Protocol (MCP) servers (both independently and within the OAI Apps SDK Framework).

-LLM integration extends to full pipeline training on cross-encoder re-rankers (mining tuples, back-propagation, training, loss function calculation, evals and harness construction, auto-rollback mechanisms based on monitored canary signals).

- Embedding Models: I utilize a diverse set of embedding models, including OpenAI, Cohere, Voyage, BGE, and sentence-transformers. I am also proactively exploring and implementing cutting-edge

models like Mamba-like S4/SSM embeddings and re-rankers, as I believe State Space Models and Test-Time Training represent the post-transformer future.

Languages & Frameworks

- Languages: (Py)Torch, Python, TypeScript, JavaScript, SQL, Bash, Go (exporters/glue)
- The backend technologies include: OpenAPI, FastAPI, Node.js/Express, Flask, along with support for asyncio and multiprocessing.
- Frontend: React, Vite, TailwindCSS, Electron, uvicorn, 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; all Linux tools, extremely proficient with linux and cmd 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 and analysis platform for clinicians and researchers interpreting qEEG brain health data.

RAG Platform for qEEG and Clinical Documentation

- 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, and full harness from LLM hallucination, with QC gates in every stage
- Implemented content-aware chunking so brain region summaries, symptom clusters, and protocol descriptions remain semantically coherent

Multi-LLM Orchestration (qEEG Council)

- 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: PDF pages rendered 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

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

---

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.

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)
- Traits-first architecture with feature gating—UI and API dynamically adapt to active provider capabilities
- Canonical event model with normalized inbound/outbound events, status mapping, standard error codes

Security & Compliance

- HIPAA-aligned: HMAC webhook signature verification, short-TTL API tokens, compliant logging architecture
- Zero PHI in logs—only document IDs and metadata surfaced for debugging while maintaining compliance
- Multi-key API authentication with scoped permissions, HTTPS enforcement, rate limiting
- GUI-first Admin Console served by API—zero-configuration UI for provider management and diagnostics
- Identical SDKs for Node.js and Python published to npm and pip with comprehensive examples
- Three MCP servers (stdio, HTTP, SSE) enabling natural language fax operations with Claude Code

Key Achievements

- Reduced webhook processing latency by 40% through async/await optimization and connection pooling
- Implemented automatic provider failover with circuit breaker pattern for 99.9% uptime
- 85% code coverage including integration tests for all supported providers

---

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

Pipeline & Retrieval Architecture

- Chunking and embedding strategies tailored for resumes and job descriptions—key signals (skills, industries, seniority, locations) remain intact
- Query pipelines combining keyword filters, vector search, and post-processing to rank candidates for recruiters
- OpenAI embeddings integrated with custom CRM platform via REST and GraphQL APIs
- AWS S3 integration for scalable document storage—resumes, job descriptions, and parsed artifacts with lifecycle policies and versioning

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 vs historical data

Leadership

- Managed 12-person recruiting team: hiring, coaching, performance management, process improvement
- Recognized as top performer among 800 recruiters—systems built were actually usable and valuable to 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

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)