Cayman Roden

AI/LLM Application Developer | RAG Pipelines · Multi-Agent Systems · MCP · Agentic AI

310-982-0492 | caymanroden@gmail.com | linkedin.com/in/caymanroden | chunkytortoise.github.io | github.com/ChunkyTortoise | Remote, US-based

PROFESSIONAL SUMMARY

AI/LLM Application Developer with a track record of shipping production AI systems end-to-end. Delivered a live Claude-powered lead qualification platform (500+ leads, zero downtime, Jan-Mar 2026). Built production RAG pipeline with 94.6% retrieval accuracy on 28-fixture adversarial eval suite. Published `mcp-server-toolkit` to PyPI — 9 pre-built MCP servers with A2A adapter. Submitted PRs to LiteLLM (27K+ stars), FastAPI (80K+ stars), and pgvector-python. 9,956+ automated tests across production repos. Seeking AI/LLM Developer, MLOps, and Agentic AI roles.

Live: docextract-demo.streamlit.app | PyPI: mcp-server-toolkit

TECHNICAL SKILLS

LLM and GenAI: Claude API (tool_use, streaming, multi-turn) | OpenAI API | Gemini API | Hugging Face Transformers | PyTorch | LangGraph | LangChain | LlamaIndex | Prompt Engineering | Chain-of-Thought | Fine-Tuning (QLoRA, LoRA, PEFT) | Model Evaluation | RAGAS | LLM-as-Judge | NLP **RAG and Retrieval:** pgvector | Pinecone | Chroma | Weaviate | BM25 | Reciprocal Rank Fusion | Agentic RAG | ReAct Reasoning | Semantic Caching | Document Chunking | Sentence Transformers | Multi-Document Synthesis **Agentic AI & MCP:** Agentic AI | AI Agent | Multi-Agent Orchestration | Model Context Protocol (MCP) | MCP Server | A2A Adapter | Tool Use / Function Calling | Function Calling | Agent State Management | Circuit Breakers | Per-Agent Model Routing | LangGraph **MLOps:** MLflow | Weights & Biases | Model Registry | A/B Testing | Model Monitoring | Experiment Tracking **Backend:** Python (asyncio, FastAPI, pytest, Pydantic, SQLAlchemy) | TypeScript | SQL | PostgreSQL | Redis | Docker | Kubernetes **Infrastructure:** AWS (EC2, S3, RDS, ElastiCache, SageMaker, Lambda) | Terraform | Render | Vercel | GitHub Actions CI/CD

OPEN SOURCE CONTRIBUTIONS

- **LiteLLM** (BerriAI/litellm, 27K+ stars): Open PR #24551 - typed exception mapping for Router fallback, surfaces AuthenticationError, RateLimitError, NotFoundError distinctly instead of swallowing as generic Exception
- **FastAPI** (80K+ stars): Open PR #15217 - BackgroundTasks interaction warning for dependency injection chains
- **pgvector-python**: Open PR #151 - async SQLAlchemy + HNSW index documentation covering connection pooling and performance-critical index selection

EXPERIENCE

AI Developer (Contract) - Remote - June 2022 - Present

- Eliminated manual lead qualification for real estate client by building 3 Claude-powered SMS bots: 500+ leads processed, under 500ms response time, bilingual EN/ES, zero downtime over 3-month production run. FastAPI, Redis, Claude API, GoHighLevel CRM. 1,702 tests.

- Built production RAG pipeline (DocExtract): 94.6% extraction accuracy on 28-fixture golden eval (12 adversarial cases including 4 prompt injection attacks). Agentic RAG with ReAct loop, semantic caching (88% hit rate), circuit breaker model fallback, RAGAS evaluation + LLM-as-judge CI gate. 1,185 tests. Live on Render.
- Architected 3-tier cache (L1 memory, L2 Redis, L3 PostgreSQL) for multi-agent orchestration platform achieving 88% aggregate hit rate. Domain-specific agent mesh (Lead Intake, Buyer, Seller) with 8 agent capabilities, circuit-breaker failover, per-agent model routing (Haiku/Sonnet/Opus), and human handoff. 6,657 tests.
- Published mcp-server-toolkit to PyPI: 9 pre-built MCP servers with A2A adapter, reducing LLM tool integration from days to a single import. 412 tests, 88% coverage.
- Built real-time voice AI pipeline: Silero-ONNX VAD, Deepgram STT/TTS, Claude reasoning, under 300ms TTFB, barge-in support. 171 tests.

Previous Career | Operations Management - 2012 - 2022

10 years client-facing operations; stakeholder communication and scope management applied to technical delivery.

PROJECTS

DocExtract AI - Production RAG Pipeline

FastAPI - ARQ - pgvector - Claude API - Streamlit

Async document processing with hybrid retrieval (BM25 + cosine + RRF), citation-aware answers, agentic ReAct reasoning, semantic caching, fine-tuning data pipeline (DPO pair generation from HITL corrections; QLoRA training infrastructure ready). Kubernetes manifests and AWS Terraform IaC (deployment-ready). 1,185 tests, 87%+ coverage. Live: docextract-demo.streamlit.app

EnterpriseHub - Multi-Agent Orchestration

FastAPI - PostgreSQL - Redis - LangGraph - Claude API - Prometheus - Grafana

Domain-specific agent mesh (Lead Intake, Buyer, Seller) with 88% cache hit rate via 3-tier cache (L1 memory, L2 Redis, L3 PostgreSQL). 8 agent capabilities, circuit-breaker failover. OWASP-hardened security (input validation, Ed25519 webhook verification, Redis rate limiting). OpenTelemetry instrumentation. 9-panel Grafana dashboard configs. 6,657 tests.

mcp-server-toolkit - Published PyPI Package

Python - MCP SDK - OpenTelemetry

9 pre-built MCP servers with A2A adapter, auto-caching, rate limiting, auth middleware. MCPTestClient for testing without live API keys. 412 tests, 88% coverage. pip install mcp-server-toolkit

CERTIFICATIONS

- IBM Generative AI Engineering with PyTorch, LangChain & Hugging Face Professional Certificate — 144h
- DeepLearning.AI Deep Learning Specialization — 120h
- Microsoft AI & ML Engineering Professional Certificate — 75h
- Duke University Large Language Model Operations (LLMOps) Specialization — 48h
- IBM RAG and Agentic AI: Build Next-Gen AI Systems Professional Certificate — 24h
- Google Cloud Generative AI Leader Certificate — 25h
- Claude Code in Action — Anthropic — 3h

21 total certifications · 1,831 hours · Google, IBM, Microsoft, DeepLearning.AI, Duke, Vanderbilt, Anthropic