

Colin Peters

Senior AI Engineer · LLM & Agentic Systems · Healthcare

Louisville, KY · (773) 466-4989 · Colin.Peters0410@outlook.com · linkedin.com/in/colin-peters-7984403b6

PROFESSIONAL SUMMARY

AI engineer specializing in LLM systems for domains where being wrong has consequences. 7+ years shipping production software and ML in healthcare, the last four building LLM and agentic systems — clinical decision support used by 1,000+ physicians, and autonomous patient intake deployed across 50+ hospitals and clinics. Depth in agent orchestration, retrieval grounding, evals and guardrails, and HIPAA-compliant deployment on AWS and Azure.

CORE SKILLS

Generative AI & LLMs: LLM application development, RAG, prompt engineering, evals and guardrails, fine-tuning (LoRA, RLHF), function calling, structured output, streaming inference, and multimodal work across text, voice, and vision. Models via AWS Bedrock, Azure OpenAI, Azure AI Foundry, Vertex AI, and Hugging Face.

Agents & Retrieval: Multi-agent orchestration (LangChain, LangGraph, LlamaIndex), tool use through function calling and MCP, and memory and context management. Hybrid semantic and keyword retrieval, chunking and ingestion strategies, embedding models, and vector stores (Pinecone, pgvector, Weaviate, Azure AI Search).

Healthcare & Compliance: HIPAA-compliant architecture, HL7/FHIR, EHR/EMR workflows, clinical NLP, clinical decision support, telemedicine, IoT and BLE medical-device integration, clinical audit logging, and FDA regulatory data infrastructure.

MLOps & Observability: Model deployment, versioning, monitoring, and drift detection. Eval, experiment tracking, A/B testing, and CI/CD for ML with SageMaker, Azure ML, Vertex AI, Langfuse, and MLflow. PyTorch, scikit-learn, XGBoost, pandas, and NumPy for NLP, time-series, and anomaly detection.

Real-Time & Voice AI: LiveKit (Agent Framework, rooms, tracks, egress and ingress), WebRTC, AI voice agents (STT, TTS, LLM), SIP/PSTN, WebSocket streaming, and low-latency audio and video pipelines.

Engineering & Infrastructure: Python (FastAPI, Django), TypeScript, Node.js, React/Next.js, C#/.NET, Java (Spring Boot), and SQL. REST, GraphQL, microservices, event-driven architecture, and Kafka. AWS, Azure, and GCP with Docker, Kubernetes, Terraform, and GitHub Actions. PostgreSQL, MongoDB, Redis, Snowflake, and BigQuery.

WORK EXPERIENCE

Independent Consultant (Self-Employed) | Senior AI / Full-Stack Engineer Remote · Nov 2025 – Present

- Advise healthcare and enterprise clients on LLM system architecture, from retrieval design and agent orchestration through HIPAA-compliant deployment on AWS and Azure.
- Ship production agentic workflows on LangGraph and LlamaIndex with tool use via function calling and MCP, including multi-model routing across OpenAI, Anthropic, and Google models to balance latency, cost, and accuracy per task.
- Build the eval harnesses and guardrails clients lack — retrieval-grounding checks, structured-output validation, and regression suites over curated case sets — so a prompt or model change can be shown to have improved something rather than assumed to.
- Deploy real-time voice agents on the LiveKit Agent Framework with integrated STT, TTS, and LLM backends, paired with React and Next.js frontends streaming token-by-token output over WebSockets.
- Own the infrastructure end to end: Terraform-provisioned AWS and Azure, FastAPI model serving, containerized deployment on Docker and Kubernetes, and Langfuse observability across prompt versions, token cost, and latency.

AdviNOW Medical | Senior AI / Full-Stack Engineer Scottsdale, AZ · Dec 2023 – Oct 2025

AI-powered autonomous medical intake and clinical decision support, deployed across physician practices and hospital systems.

- Built the patient-intake LLM pipeline (Python, LangChain, OpenAI and Anthropic APIs) behind 10,000+ clinical encounters, cutting average intake time 30–50% through NLP-driven symptom extraction and structured clinical reasoning.
- Engineered stateful LangGraph agent workflows supporting branching clinical decision paths, conditional tool execution, and multi-turn context management, emitting ranked differentials, recommended labs, and medication suggestions as schema-validated output consumed directly by the frontend.
- Architected a LiveKit real-time telemedicine platform deployed across 50+ hospitals and clinics, adding AI voice-agent pre-consultation intake and IoT medical-device streaming, plus a Remote Medical Volunteer System pairing video with LangGraph triage agents for structured remote encounters.
- Grounded high-stakes clinical output in RAG over proprietary medical knowledge bases (Azure AI Search, pgvector), with Langfuse evals and HIPAA-compliant audit logging tracking prompt versions, token cost, latency, and output quality.
- Benchmarked foundation models across AWS Bedrock, Azure OpenAI, Azure AI Foundry, and Vertex AI on latency, clinical accuracy, and cost, reducing inference spend by routing routine extraction to smaller models and reserving frontier models for

diagnostic reasoning; ran SageMaker MLOps to fine-tune, version, A/B test, and monitor clinical NLP models trained on proprietary medical terminology.

- Delivered REST and GraphQL APIs (FastAPI, Node.js) with Redis caching and Kafka event streaming, plus C#/.NET and Java (Spring Boot) integration services for EHR interoperability — shipped on Kubernetes and Terraform with AES-256 encryption and RBAC, across a 5–10 engineer team.

Glass Health | Full-Stack Engineer

San Francisco, CA · Sep 2021 – Nov 2023

AI clinical decision support delivering differential diagnosis, SOAP note generation, and evidence-based care plans via LLMs to 1,000+ physicians.

- Built and iterated the core NLP and prompt-engineering framework behind differential diagnosis, SOAP note authoring, ICD-10 suggestion, and treatment planning for 1,000+ physicians on the public platform.
- Introduced multi-step clinical reasoning chains covering case decomposition, retrieval over clinical references, and structured assessment synthesis, raising diagnostic completeness against a curated internal case library.
- Developed the prompt and eval tooling that benchmarked every prompt change against clinical-accuracy metrics on real-world case sets, validated with licensed physicians and clinical advisors.
- Designed a Python/FastAPI inference microservice handling prompt construction, model orchestration, output validation, and multi-turn context — decoupled from the Node.js app for independent scaling, and streaming token-by-token to the physician UI instead of waiting on batch inference.
- Persisted case histories and outputs in PostgreSQL and MongoDB on GCP Cloud Run with Docker autoscaling, and contributed Ruby on Rails admin tooling and C#/.NET integrations.

Belun Technology | Full-Stack Engineer

Hong Kong · Dec 2018 – Aug 2021

FDA-registered wearable IoT devices for sleep apnea detection and remote patient monitoring across international clinical trials and home care.

- Built real-time IoT biometric ingestion pipelines (SpO2, heart-rate variability, accelerometry, automated sleep staging) from BLE wearables for 10,000+ monitored patients across multiple countries.
- Developed anomaly-detection algorithms over biometric time-series to flag oxygen-desaturation and arrhythmia events in real time, contributing to the device's FDA and international regulatory clearance.
- Engineered fault-tolerant AWS pipelines (S3, Lambda, RDS) with buffering, deduplication, and reliable writes for intermittent BLE connections, on PostgreSQL schemas optimized for longitudinal time-series data.
- Built React clinician dashboards surfacing overnight SpO2 trends, AHI scores, and sleep-stage hypnograms, backed by Node.js REST APIs and Python sleep-staging services, plus Ruby on Rails compliance-export pipelines used in regulatory submissions.

EDUCATION

B.S., Computer Science | City University of Hong Kong

Hong Kong · 2014 – 2018