

Gregory Chin

gregory.chin410@outlook.com | 254-836-8666 | Fort Lee, NJ 07024 | linkedin.com/in/greg-chin-6167533b4

SUMMARY

AI Full-Stack Engineer with 8+ years building LLM-powered systems, real-time voice/video platforms, and high-scale backend AI infrastructure across healthcare, fintech, and consumer AI. Deep expertise in RAG pipelines, multi-agent orchestration, and HIPAA-compliant clinical AI, backed by hands-on ML engineering (model fine-tuning, MLOps, embeddings) and applied data science. Delivered systems handling 150k+ daily video events, thousands of concurrent autonomous agents, and voice AI deployments earning 4.8/5 satisfaction scores. Polyglot engineer fluent in Python (AI/ML layer), Go (high-throughput APIs), Java/Spring Boot and C#/.NET (enterprise services), and Rust (performance-critical modules).

SKILLS

AI / LLM Engineering: RAG (dense/sparse/hybrid), multi-agent orchestration, fine-tuning (LoRA/QLoRA), prompt engineering, LLM evaluation (RAGAS, LLM-as-judge), DSPy, OpenAI GPT-4/5, Claude, Gemini, Mistral

ML Engineering / MLOps: PyTorch, scikit-learn, XGBoost/LightGBM, fine-tuning (LoRA/QLoRA), MLflow, Weights & Biases, experiment tracking, model versioning (DVC), feature stores (Feast), model monitoring & drift detection, Airflow/Prefect, SageMaker, Vertex AI, ONNX, model serving (BentoML, Triton), CI/CD for ML

Data Science / Analytics: Python (Pandas, NumPy, Polars), SQL, statistical analysis, A/B testing, EDA, feature engineering, data visualization (Matplotlib, Plotly), Jupyter, Spark/PySpark, time-series analysis, computer vision (OpenCV), NLP

Agentic Frameworks: LangChain, LangGraph, LangSmith, Langfuse, AutoGen, CrewAI, MCP, tool-calling, memory, guardrails

AI Platforms: AWS Bedrock, AWS SageMaker, Google Vertex AI, Azure AI Foundry, Azure OpenAI, Hugging Face

Vector Databases & Search: Pinecone, Qdrant, Milvus, Weaviate, pgvector, Chroma, OpenSearch, Elasticsearch, Azure AI Search, embedding strategies, reranking, hybrid search

Voice & Real-Time: LiveKit, WebRTC, Pipecat, Twilio, Deepgram, ElevenLabs, Cartesia, Whisper, VAD, FreeSWITCH

Video & Streaming: FFmpeg, OpenCV, RTSP, AWS MediaConvert, Stable Diffusion, ComfyUI, script-to-video, avatar animation

Backend, by role

- Python (AI/ML layer): FastAPI, LangChain, RAG pipelines, data engineering, model serving
- Go: high-throughput microservices, low-latency APIs, gRPC, event-driven services
- Java: Spring Boot, enterprise backend, Kafka, data ingestion services
- C# / .NET: ASP.NET Core, EHR integrations, middleware pipelines, admin tooling
- Rust: performance-critical modules, stream processing, FFI integration with Python/Node.js
- TypeScript / Node.js: NestJS, real-time APIs, event-driven microservices
- Ruby on Rails: MVP prototyping, REST APIs, background jobs (Sidekiq)

Data: PostgreSQL, Redis, pgvector, MongoDB, MySQL, Supabase, schema design, query optimization, connection pooling

Cloud & Infra: AWS (Lambda, EC2, S3, API Gateway, Bedrock, SageMaker, CloudWatch, MediaConvert), GCP, Azure, Docker, Kubernetes, Terraform, CI/CD

Industry-Specific

- Healthcare: HIPAA-aligned AI systems, PHI/PII handling, RBAC, audit trails, EHR integration, clinical NLP, FHIR-adjacent pipelines
- Fintech: lead generation automation, real-time transaction pipelines, CRM/EMS integration, webhook orchestration
- Retail / IoT: RTSP camera ingestion, real-time sensor pipelines, foot-traffic analytics, edge processing

Automation: n8n, outbound/inbound lead agents, marketing automation, CRM integration

Frontend: React, Next.js, Vue.js, Angular, React Native, iOS, API contract design and cross-platform delivery

EXPERIENCE

Self-Employed, Senior AI Software Engineer 07/2025 – Present

- Built voice AI agents for healthcare and SaaS clients using LiveKit, Pipecat, Deepgram, and ElevenLabs, reducing response time 30-40% and achieving a 4.8/5 satisfaction score.
- Designed multi-stage RAG pipelines (LangChain, LangGraph, Pinecone/Qdrant) for clinical, insurance, and knowledge retrieval tasks, improving relevance and reducing query latency.
- Built multi-agent systems with LangGraph, including orchestrator and domain agents with shared Redis memory, tool sets, and output guardrails, automating healthcare intake, insurance pre-auth, CRM/EMS entry, and appointment scheduling, reducing manual processing time ~50%.
- Fine-tuned open-source LLMs with LoRA/QLoRA on domain datasets and tracked experiments with MLflow and Weights & Biases, versioning models and datasets to keep training reproducible across iterations.
- Leveraged AWS Bedrock Data Automation for document extraction, summarization, and routing, and used LangSmith for tracing, evaluation, and prompt optimization in production.
- Deployed enterprise LLM solutions via Azure AI Foundry and Azure OpenAI, using Azure AI Search for hybrid retrieval (vector + keyword) over structured and unstructured enterprise data.
- Built AI-powered outbound/inbound lead generation agents integrated with CRMs via gRPC and webhooks, increasing qualified lead volume 30-40% and cutting manual calling effort ~50%.
- Built fintech and marketing automation pipelines (email/DM sequencing, LLM personalization) using n8n and FastAPI, lifting reply rates 20-25% and compressing campaign setup from days to hours.
- Delivered Java Spring Boot microservices and C# ASP.NET Core APIs for enterprise client integrations, exposing secure versioned REST endpoints consumed by web and mobile frontends.
- Authored performance-critical Rust modules for low-latency data transformation and stream processing, integrating compiled Rust components via FFI into Python and Node.js service boundaries.
- Built Ruby on Rails backends for rapid MVP engagements using ActiveRecord, Sidekiq, and RESTful conventions.
- At Glass Health: automated clinical documentation by integrating EHR systems with LLM pipelines, fine-tuned LLMs on clinical datasets using LoRA, and delivered versioned model API backends.
- At Ringfree: built multi-agent orchestration coordinating qualification, scheduling, and follow-up agents with shared Redis memory, and built a real-time transcription pipeline extracting intents and syncing summaries to CRM.

Butterflies AI, Senior Software Engineer 12/2023 – 06/2025 | Seattle, WA

- Led engineering for a 7-person team building an AI social network on AWS Bedrock and LangGraph, supporting millions of daily users across web, iOS, and React Native.

- Architected autonomous AI agent pipelines for thousands of concurrent characters, handling multi-turn memory, persona consistency, safety filtering, and prompt orchestration, and added LLM-as-judge evaluation that reduced moderation incidents.
- Built generative character creation services using ComfyUI, Stable Diffusion, and custom LoRA training pipelines, enabling AI persona creation in minutes and driving thousands of new characters per week.
- Designed event-driven microservices (message queues, priority workers, Pub/Sub) for image generation, feed updates, and notification delivery, sustaining sub-second feed latency at millions of AI-generated events per day.
- Implemented scalable Node.js and Supabase backend (row-level security, auth, rate limiting) on AWS, shipping features that grew AI-generated content 50%+ quarter-over-quarter and improved user retention.
- Wrote Java-based data ingestion services and C# .NET utility tooling for internal analytics pipelines and cross-team integrations across a polyglot engineering org.

Theoria Medical, Software Engineer 02/2019 – 11/2023 | Novi, MI

- Led a 4-person platform team building AI systems processing thousands of patient encounters monthly under HIPAA and SOC2 requirements.
- Built production clinical document processing pipelines using LLM and vision APIs, ingesting intake forms, insurance documents, and clinical notes into pgvector and Pinecone with structured chunking, embedding, and retrieval.
- Developed a real-time AI medical receptionist (Twilio + ASR/TTS + LLM) handling inbound patient calls end-to-end: identity verification, intent capture, triage scoring, and smart routing, measurably reducing call-center volume and handle time.
- Engineered pre-visit AI pipelines (voice + form ingestion, LLM summarization, structured SOAP notes) saving several minutes per clinical encounter.
- Designed HIPAA-aligned AI microservices in FastAPI with RBAC, PHI/PII isolation, environment segregation, and audit logging, integrating with web apps, telephony, and EMS systems.
- Built LLM-powered patient-communication automation (reactivation campaigns, reminders, educational content) using Python, FastAPI, and AWS, eliminating manual drafting for the operations team.
- Developed ASP.NET Core (C#) services for EHR vendor integrations and internal admin tooling, enforcing data validation and HIPAA-compliant audit trails.

Dayta AI, Software Engineer 08/2017 – 01/2019 | Hong Kong

- Built real-time IoT data pipelines for the Cyclops retail intelligence platform, ingesting RTSP camera streams, processing 150k+ daily videos, and integrating POS and sensor data for foot-traffic heatmaps and peak-hour alerts.
- Developed CV-powered demographic tracking (age/gender via OpenCV) and anonymized RTSP stream processing with role-based dashboards for retail and mall operations.
- Created script-to-video pipelines transforming raw camera data into narrated shopper-behavior reports using FFmpeg and TTS, eliminating manual video editing for marketing teams.
- Implemented cloud-rendered BI report pipelines (AWS MediaConvert + Three.js) reducing analytics delivery time ~60% for global retail clients.
- Built Ruby on Rails REST APIs and Sidekiq background workers powering the Cyclops dashboard's data ingestion and alert-delivery layer across multi-tenant deployments.

EDUCATION

B.S. Computer Science | Hong Kong College of Technology (HKCT) | 2013–2017