Johnson Subedi

johnsonsubedi.com subede.johnson@gmail.com +1 (737) 297-8567 github.com/i-johnson

Experience

Software Engineer 06/2025 – Present

Coalition San Francisco, CA

- Developed **Quotes-MCP** (schema design → containerized deploy → org-wide rollout): exposing quote data as **30+** agent-consumable tools, adopted by **50+ engineers & PMs** and cutting product-support tickets **35%** (**40** → **25/wk**)
- Built Pitch Genius end-to-end (pipeline → LLM synthesis → rendering → rollout): a fault-tolerant engine fusing claims, coverage, and client-need data into broker-branded presentations, scaling to **3K+** decks and **3 hrs** → **2 min** prep, with p95 held flat under **5**× load via bounded-concurrency offloading and timeouts
- Built the broker email intake pipeline as an **event-driven ingestion system** (inbound email events → 8-class classifier → Temporal orchestration → attachment parsing → submission prefill): reclaiming **∼35 hrs/day** (**≈4 FTE**) of manual triage; made retries safe with Message-ID idempotency keys and S3 attachment offload
- Co-led the short insurance application end-to-end (config schema → dynamic document rendering → cross-team rollout): a config-driven form engine for **batch document generation**, cutting applicant fields (**35** → **10**), processing **22M+** applications across **7 geos** with generated documents persisted to S3, driving **25% quoting growth**
- **Tech stack**: *Flask, Go, AWS, Temporal, LLM tooling (MCP), Datadog*

Software Engineering Intern 05/2024 – 08/2024

Genentech San Francisco, CA

- Built the clinical trial concept management app (Postgres + trigger versioning → 14-tool LangChain agent): replaced **600+ Google Sheets** with a system of record whose agent read and mutated protocol concept tables through Pydantic-validated tools, versioning every mutation as a revertible revision; **96%** tool-selection accuracy
- Architected and deployed **multi-modal RAG** pipeline over clinical trial protocols using Milvus (HNSW indexing) and NLTK-optimized semantic chunking; tuned HNSW params against recall/latency/memory tradeoffs with a labeled query set, recall@k baselines, and repeated runs per config; improved retrieval accuracy **45%** and cut protocol analysis from 40 to 14 hrs/week
- **Tech stack**: *FastAPI, PostgreSQL, LangChain, TypeScript, Milvus, DynamoDB*

Software and Data Engineering Intern Co-Op 08/2023 – 05/2024

Valvoline (R&D) Lexington, KY

- Built and shipped an **ML pipeline** predicting tire Remaining Useful Life for **300K+ vehicles**: streamed **3M+ sensor data points** from 5 sources through **Kafka**, batch-processed with parallel **PySpark** orchestrated by **Airflow** and landed into S3, engineered features and trained with cross-validation to **98% accuracy**, and deployed for live inference in production
- Expanded appointment scheduling module using Spring Boot and Apache Kafka for real-time task queuing
- **Tech stack**: *Python, Spring Boot, Airflow, PySpark, TypeScript*

Education

BSc in Computer Science and Mathematics, Centre College 08/2022 – 05/2025

- **Coursework:** Data Structures & Algorithms, Software Design, Database Systems, Computer Networks, Systems Programming

Projects

ContractorOps.ai | *FastAPI, Next.js, Twilio, AWS (S3, Bedrock, Transcribe)* 11/2025 – Present

- Founded (**0→1**): an AI back-office for contractors with **30 paid customers**; **300+** discovery calls shaped the roadmap
- NovaSonic voice on Twilio calls, a lead-gen agent over Google search, and agentic workflows that take clients from leads to payment

Real-time Spanish Conversation Tutor | *vLLM, PyTorch/PEFT, Core ML, K8s, Prometheus* 05/2026 – Present

- Self-hosted Whisper + **7B** vLLM stack (vs. hosted ASR + 70B); **64**-way concurrency via batching & prefix/KV-cache tuning
- Distilled **70B→7B** (synth data → LoRA SFT → 4-bit) at teacher parity on a **100**-sample eval; on-device ASR via Core ML

Sphere | *Express, pgVector, Redis, BullMQ, React* 11/2024 – 04/2025

- Automated trading engine placing live Alpaca orders; Redis & BullMQ for job scheduling, concurrency control, and state persistence
- Custom parsing & hierarchical semantic chunking for **RAG**; pgVector & MongoDB for retrieval over **300+ page** SEC filings

Skills and Interests

Focus: Inference serving, parallel and event-driven pipelines, agent orchestration and evals; active in **GPU MODE** (CUDA/Triton)

Languages & Frameworks: Go, Python, Java, JavaScript, TypeScript, C++, Django, Flask, FastAPI, Express, React, Next.js

AI/ML: PyTorch, LangChain, RAG, MCP, AWS Bedrock, Milvus, pgVector

Technologies: AWS (Lambda, SQS, Kinesis), Kafka, Spark, Airflow, Docker, Kubernetes, Git, PostgreSQL, DynamoDB, Redis