

Corey Hall

Address: Katy, TX

Email: coreyhall84840@outlook.com

Phone: (435) 270-0555

LinkedIn: <https://www.linkedin.com/in/coreyh9390>

Summary

Senior AI Software Engineer with 9+ years of experience designing and delivering enterprise-scale AI and full-stack solutions across healthcare, insurance, and real estate in both startup and enterprise environments. Expert in Generative AI, Agentic AI, Retrieval-Augmented Generation (RAG), AI automation workflows, computer vision, NLP, LLM applications, and speech-to-text/text-to-speech systems. Leads end-to-end development of generative and agentic AI systems, delivering RAG pipelines and real-time voice AI that cut manual effort by 40%+ and serve thousands of users. Partners with founders and product teams to scale cloud-native platforms and measurable business outcomes.

Skills

- **Programming Languages:** Python, Java, JavaScript/TypeScript, SQL
- **AI & ML:** Generative AI, Agentic AI, RAG Pipelines, LLM Frameworks (LangChain, LangGraph), Prompt Engineering, NLP, Computer Vision, Machine Learning
- **Voice & Real-Time AI:** Real-Time STT/TTS Pipelines, Streaming Inference
- **Cloud AI/ML Platforms:** AWS (SageMaker, Bedrock, Lambda AI Workflows), GCP (Vertex AI, Pub/Sub), Azure (Cognitive Services, OpenAI on Azure)
- **Backend Development:** Node.js, Python (Django, FastAPI, Flask), Java (Spring Boot, J2EE, Hibernate), Event-Driven Microservices, Serverless Architectures
- **Frontend Development:** React, Next.js, Angular
- **Cloud & DevOps:** Docker, Kubernetes, Terraform, CI/CD, Cloud Platforms (AWS, Azure, GCP), Git, Version Control
- **Databases:** PostgreSQL, MongoDB, DynamoDB, Redis, Vector Stores (Pinecone, FAISS)

Work Experience

Senior AI Full Stack Engineer

09/2023 ~ 08/2026

Aeries Technology

- Architected and delivered Generative AI and Agentic AI solutions for healthcare organizations, automating clinical documentation and administrative workflows and reducing manual effort by over 40%.
- Led architecture and development of a real-time AI voice communication platform for healthcare providers using React, Python/Django, WebRTC, SIP, and FreeSWITCH, supporting over 10,000 concurrent sessions with sub-1-second latency.
- Engineered AI-powered voice routing, transcription, and automated call orchestration workflows, improving intelligent call distribution accuracy by ~45% and streamlining patient support operations.
- Built advanced Retrieval-Augmented Generation (RAG) architectures that integrated DynamoDB vector stores, LangChain knowledge retrieval, and tool-calling agents, delivering context-aware clinical assistance that reduced average decision time for clinicians.
- Developed Python microservices with FastAPI on AWS serverless infrastructure, providing reliable and secure support for healthcare applications and achieving continuous high-availability compliance.
- Partnered with executive stakeholders, clinicians, and product leaders to define AI strategy and identify automation opportunities using Python, LangChain, and Terraform, resulting in streamlined workflows and measurable cost savings.

AI Full Stack Developer

04/2020 ~ 08/2023

Shimento Inc

- Spearheaded development of AI-powered insurance intelligence platforms that automated document processing, policy analysis, claims triage, and risk assessment, reducing manual review efforts by 45%.
- Built OCR, computer-vision, and NLP pipelines with Tesseract, OpenCV, and spaCy that processed over 1 million insurance documents annually, improving extraction accuracy.
- Designed and deployed RAG-based AI systems leveraging LLMs, vector databases, and semantic retrieval, improving information discovery efficiency by 70%.
- Led full-stack architecture using React, Next.js, Python, FastAPI, AWS, Docker, and Kubernetes, delivering scalable cloud-native AI products that supported the company's growing client base.
- Established engineering standards for prompt engineering, LLM evaluation, AI observability, and production deployment, accelerating adoption of Generative AI across the organization.
- Mentored engineers and collaborated with leadership to define AI roadmaps and technical strategy, applying Python, Azure Cognitive Services, and AWS SageMaker, which clarified project priorities and accelerated delivery of AI features.

Full Stack Developer

08/2017 ~ 03/2020

Thoughtgreen Technologies Inc

- Designed scalable data pipelines using Python and Docker that integrated POS, inventory, and supply-chain systems, processing tens of millions of transaction records each year and delivering faster data access for analytics teams.
- Developed real-time dashboards and reporting platforms that reduced manual reporting efforts and improved operational visibility across regional teams.
- Implemented predictive analytics and demand forecasting capabilities that contributed to a 15% reduction in inventory shortages and improved stock planning accuracy.
- Modernized legacy applications by adopting an API-first design, automated testing, and CI/CD pipelines in GitLab, which shortened deployment cycles and enabled more frequent releases.
- Partnered with business stakeholders and operations leaders to build technology solutions using streaming inference and computer-vision tools, resulting in higher inventory accuracy, smoother operations, and better customer experience.

Education

Texas A&M University

Bachelor's Degree · Computer Science (2011 - 2016)