

Brent Oakes

Senior AI Full Stack Engineer

brentoakes025@gmail.com | (432) 322-4207 | Chicago, IL | [linkedin.com/in/bren-oakes-5447163b3](https://www.linkedin.com/in/bren-oakes-5447163b3)

SUMMARY

Senior AI Full Stack Engineer with 8+ years of software engineering experience, including 4+ years building production AI applications and enterprise AI platforms. Experienced in large-scale RAG systems, agentic workflows, multi-model inference, LLM evaluation, and cloud-native architectures using Python, FastAPI, TypeScript, and AWS. Built AI platforms spanning more than 1 million documents and supporting 5,000+ concurrent users, with measurable improvements in response quality, latency, and operational efficiency.

SKILLS

Applied AI and Generative AI — LLM Applications, LangChain, LangGraph, Retrieval-Augmented Generation, Agentic Workflows, AI Agents, Multi-Agent Systems, Tool Calling, Model Context Protocol, Conversational AI, Voice AI, Semantic Search, Hybrid Retrieval, Embedding Pipelines, Context Engineering, Model Routing, Prompt Optimization

AI Quality, Safety, and Operations — LLM Evaluation, Groundedness Testing, Hallucination Mitigation, Guardrails, Prompt Regression Testing, Retrieval Evaluation, A/B Testing, Model Fallback Strategies, Production Monitoring, Inference and Cost Optimization

Languages & Full-Stack Engineering — Python, TypeScript, JavaScript, SQL, FastAPI, Pydantic, Node.js, NestJS, React, Next.js, REST APIs, GraphQL, WebSockets, Asynchronous Processing, Streaming APIs

Cloud & Platform Engineering — AWS Bedrock, ECS, EKS, EC2, Lambda, S3, RDS, Docker, Kubernetes, Terraform, CI/CD, Infrastructure as Code, Autoscaling, Serverless Architecture, Azure, GCP

Data & Retrieval — PostgreSQL, SQL Server, MySQL, MongoDB, DynamoDB, Redis, Pinecone, FAISS, Neo4j, Vector Search, Hybrid Search, Data Modeling, Indexing, Query Optimization

Architecture & Security — Distributed Systems, Microservices, Event-Driven Architecture, High-Availability Systems, API Security, OAuth 2.0, JWT, Role-Based Access Control, Auditability, Performance Optimization, HIPAA-Aligned Application Controls

PROFESSIONAL EXPERIENCE

Senior AI Software Engineer, Elastech

07/2023 – 06/2026 | Arlington Heights, IL

- Led the architecture and production launch of a clinical decision-support platform that enabled healthcare professionals to retrieve treatment guidance, patient insights, and operational knowledge through natural-language interactions, accelerating access to critical information.
- Architected a large-scale RAG platform spanning more than 1 million clinical and enterprise documents, combining semantic and keyword retrieval, metadata filtering, reranking, vector search, and contextual grounding to reduce unsupported responses by 22% in internal evaluations.
- Built agentic workflows with LangGraph, LangChain, and Model Context Protocol that enabled multi-step reasoning, controlled tool execution, contextual memory, and workflow automation across clinical and operational use cases, streamlining processes and reducing manual effort.
- Engineered asynchronous Python and FastAPI inference services with streaming, batching, caching, and autoscaling, supporting more than 5,000 concurrent users while reducing end-to-end response latency by 35%.
- Integrated OpenAI, Anthropic Claude, Google Gemini, AWS Bedrock, and open-source LLMs through a model-agnostic inference layer with policy-based routing, fallback strategies, A/B evaluation, and production monitoring, improving availability and controlling inference costs.

- Established automated AI quality gates within CI/CD pipelines for prompt regression testing, retrieval quality, groundedness, hallucination risk, safety validation, and model-version comparison, increasing confidence in production releases.
- Partnered with clinical, product, security, and engineering stakeholders to translate healthcare requirements into scalable, HIPAA-aligned AI services while setting technical direction for backend architecture, cloud infrastructure, data access, and production reliability.

AI Full Stack Engineer, Grayphite

05/2021 – 06/2023 | San Jose, CA

- Delivered AI-powered customer-engagement capabilities for a real estate SaaS platform, including conversational property search and automated listing generation, helping users discover relevant properties and create market-ready content more efficiently.
- Architected end-to-end RAG workflows using Python, FastAPI, Node.js, React, and vector search, improving the relevance and contextual accuracy of customer-facing search and support experiences.
- Integrated speech-to-text, LLM reasoning, and text-to-speech services into real-time conversational and voice AI workflows, reducing perceived interaction latency by approximately 32% and creating a more natural user experience.
- Created persistent conversation-memory services using PostgreSQL, DynamoDB, Redis, and vector embeddings, supporting contextual personalization, reliable multi-turn interactions, and long-running sessions across devices.
- Deployed cloud-native AI services on AWS using Docker, ECS, Lambda, Terraform, and CI/CD pipelines, strengthening deployment consistency, scalability, and operational reliability.
- Collaborated with product managers, designers, and backend engineers to identify high-value AI opportunities, translate customer and business requirements into technical designs, and integrate AI capabilities into existing SaaS workflows.

Full Stack Developer, Epic

06/2018 – 04/2021 | Madison, WI

- Developed enterprise web applications, workflow tools, and analytics dashboards using React, JavaScript, Node.js, and Python, improving operational efficiency and data visibility for healthcare teams.
- Designed REST APIs and microservices supporting reporting, healthcare analytics, and workflow automation, enabling dependable daily processing across distributed application services.
- Implemented authentication and authorization services using OAuth 2.0, JWT, and role-based access controls, strengthening protections for applications handling sensitive healthcare information.
- Optimized PostgreSQL and MySQL performance through schema improvements, indexing, query-plan analysis, and query tuning, reducing execution times by approximately 30% and improving application responsiveness.
- Modernized AWS-hosted applications using EC2, ECS, CloudFormation, and CodePipeline, automating deployments and improving system uptime by 15% while strengthening distributed-systems and cloud-engineering practices.

EDUCATION

Bachelor of Science, Trinity University

08/2014 – 05/2018 | San Antonio, TX