**Bobby Estes**

Senior AI/ML Backend Engineer, System Designer

**Email:** bobbyestes.ai@gmail.com, **Phone:** +18632130426, **Location:** Florida, United States, **LinkedIn:** https://bobby-linkedin-redirect.vercel.app/, **Website:** https://bobby-estes.vercel.app/

## SUMMARY

## SKILLS

## PROFESSIONAL EXPERIENCE

### InsoftAI - Senior Machine Learning Engineer

02/2023 - present

* Defined and drove the long-term ML technical strategy in alignment with product and business goals, ensuring the development of AI-driven platforms capable of handling up to 90% of customer inquiries.
* Acted as a thought partner to founders on AI strategy and vision, leading the development of Support-nGen™ and LLM Twin, enhancing customer service and brand creation.
* Led, grew, and mentored a high-performing ML team, fostering a culture where people feel empowered, supported, and technically challenged.
* Ensured strong cross-functional collaboration with Product and Engineering teams, creating a clear, focused ML roadmap with visibility across functions.
* Helped get state-of-the-art models into production quickly, implementing best MLOps & LLMOps practices like experiment trackers, model registries, and prompt monitoring.
* Integrated vector databases to store knowledge efficiently, building knowledge bases from respected sources.
* Scaled the team intentionally through smart hiring, clear roadmaps, and a low-meeting culture, resulting in significant improvement in experimentation throughput and continued excellence in GenAI and computer vision research and delivery.

### Brainhub - AI/MLOps Engineer

10/2019 - 12/2022

* Developed Sierra.ai, revolutionizing document management and information accessibility, resulting in a 30% increase in operational efficiency for clients.
* Created a real-time personalized recommender system for H&M fashion articles, leveraging a two-tower model design architecture and Hopsworks AI Lakehouse.
* Led the formulation and execution of technical strategies that align with business goals, contributing to a 25% growth in user adoption rates.
* Designed and implemented three core ML serving architectures, balancing trade-offs between low latency and high throughput to optimize user experience.
* Adopted MLOps best practices, including Infrastructure as Code (IaC), CI/CD, monitoring, experiment tracking, and model registries, ensuring the system is reproducible, testable, and trackable.
* Deployed real-time recommendations using Hopsworks Serverless and KServe, resulting in improved model performance and responsiveness in production.

### Kensho - Backend-heavy AI Developer

09/2016 - 09/2019

* Built an inference pipeline in LangChain as a serverless RESTful API, enabling real-time financial question answering using RAG/TAG.
* Extended Meta’s Llama 3 model with a multimodal projector, allowing direct audio input for faster responses compared to traditional ASR-LLM combinations.
* Developed a serverless continuous training solution that fine-tunes an LLM on financial data, optimizing model performance through automatic tracking and registry saving.

### Dana Scott Design - Full Stack Developer/Assistant

02/2014 - 08/2016

* Translated Figma designs into user-friendly, reusable React components, optimizing communication and dataflow for enhanced application functionality.

## EDUCATION