Hello,

I am Bobby.

<https://bobby-estes.vercel.app/>

After carefully reviewing the company's job description, I think my skillset and hands-on experience would be the best fit for it and there would be more opportunities that I can participate in large-scale and challengable projects.

As a senior AI/ML Full stack engineer, I have been working for 10 years to develop AI-driven applications and system design architecture to ensure streaming and consistency between distributed components within the ML system.

I specialize in architecting, designing, and implementing cutting-edge software solutions tailored to meet diverse business requirements.

As a Tech Lead in various IT companies, I have successfully guided teams in developing innovative AI-driven software across healthcare, finance, and marketing. I led and developed powerful AI-driven platforms and ML projects, streaming business operations by integrating AI

driven systems capable of handling up to 90% of customer inquiries.

I solved challenges in developing ML systems and deploying ML models with advanced tech. I implemented MLOps/DevOps best practices with AWS, GCP, and Azure services across the financial, e-commerce(personalized ranking system), healthcare, marketing, and finance.

I built and deployed a real-time personalized recommender for H&M fashion articles using the 4-stage recommender architecture, the two-tower model design, and the Hopsworks AI Lakehouse.

Developed Finance analyzer/advisor platform using multi-AI agents(+voice agent) with Langchain for trading analyzer and fine-tuned LLM using Tensorflow, pytorch, and scikit-learn.

Especially, implemented advanced 3 pipeline ML architecture(Feature/Training/Inference), advanced chunking strategy, advanced RAG/TAG techniques, FEA architecture in Graph RAG, and Router system that can dynamically choose the system between semantic search, Graph RAG, and Elastic search.

Also implemented binary quantization tech for improving Rag search to 40x faster!

Also, I optimized and scaled ML systems involving large language models, implementing the ideas on centralizing LLM functions, reducing variation, and improving error handling, failover mechanisms, and best approaches like creating a function server for native code execution and enhancing client libraries for better performance and standardization.

And designed a 4-stage architecture to build a system that can handle recommendations from a catalog of millions of items and a two-tower model, a flexible neural network design that creates embeddings for users and items.

I implemented a strategy similar to what TikTok employs for short videos, which will be applied to H&M retail items, and designed and implemented a highly scalable and modular Python package that orchestrates the ML steps into 3 fully automated advanced 3 ML pipeline - Feature, Training and Inference pipeline to solve all challenges remains in ML systems such as inconsistency between distributed components, I/O bottleneck on the database, utilized CDC patterns and bytewax to enhance system’s latency and build event-driven architecture.

Especially, I built a tabular semantic search retrieval service instead of Text-to-Sql by llm for multi-attribute structured data(prices, ratings, categories) beyond just text for Amazon e-commerce products.(SuperLink + MongoDB)

And I improved the iteration speed within the team by automating MLOps practices such as Structured Model Experimenting, Resources Monitoring, Model Registry, CI/CD, and Jira Integrations and leveraging the FTI architecture on top of Hopsworks AI Lakehouse.

Also, I have hands-on experience with developing AI agents with Langchain, Llama-index, downstream tasks such as developing Multimodal LLM, computer vision tasks, and Fine-tuning LLM using Tensorflow, python, and sci-kit-learn and Python backend framework.

With a deep understanding of Transformer architecture and speech recognition, I developed a fast, and extensible multimodal LLM that can understand text as well as human speech, without the need for a separate Audio Speech Recognition (ASR) stage. In this project, after research like AudioLM, SeamlessM4T, Gazelle, and SpeechGPTs, I have extended Meta's Llama 3 model with a multimodal projector that converts audio directly into the high-dimensional space used by Llama 3. This direct coupling allows this to respond much more quickly than systems that combine separate ASR and LLM components.

Innovatively, this multi-modal LLM can natively understand the paralinguistic cues of timing and emotion that are omnipresent in human speech.

Also, I focused on working with the advanced RAG system to improve the accuracy rate to 97% by integrating various techniques at each step using Euclidean, cosine similarity, and advanced RAG/TAG techniques.

I thoroughly monitor A/B testing and evaluate the experiments using Cuda-enabled Nvidia A100, Comet ML’s experiment tracker, and save the best model to Comet’s model registry. (deployed on Qwak, AWS) and deploy it as a REST API on Qwak.

Additionally, have extensive experience in managing and maintaining server infrastructure, cloud architecture, and automation for high-performance AI applications, specializing in building microservices and cloud-based solutions on AWS and Azure.

Right now, I am looking for a new opportunity.

Let's discuss this in more detail.