

Bobby Estes

Senior Software, Gen AI/ML Backend Engineer, System Designer

Florida, United State(US citizen)
bobbyestes.ai@gmail.com
<https://bobby-linkedln-redirect.vercel.app/>
<https://bobby-estes.vercel.app/>
+18632130426

I am an ambitious Senior AI Engineer, with 10 years of experience in the Software Development industry.

My passion for Artificial Intelligence research and development ignited at its very beginning in America. Since then, I've been keen on architecting, designing, and implementing top-of-the-line software solutions tailored to the unique needs of businesses. My commitment to staying at the forefront of technological advancements has enabled me to exceed the evolving demands of the digital business landscape.

My biggest differentiator is my expertise - based upon best practices study, a non-conventional approach that goes beyond the latest tech trends, and proven solutions that best fit business objectives. Whether we're talking about Product Development, driving projects as a Contractor, I'm enthusiastic about delivering results that transcend expectations.

My proficiency in AI, MLOps, and System Architecture are not just skill sets. They are components that bridge the gap between real-world solutions and advanced algorithmic strategies.

- Technical Proficiencies

Languages:	Python, JavaScript/TypeScript, C++
Software Design Architecture:	FTI Architecture, batch serving architecture, online real-time pipeline, offline batch pipeline, asynchronous inference pipeline
AI Frameworks:	PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, langchain, Llama-index, Haystack, langGraph, AutoGen, Crew AI, Agentic Transformer
MLOps	GCP(Vertex AI, GCR, GKE, GCS, pub/sub), AWS(AWS SageMaker, Fargate, lambda, S3 bucket), Azure, W&B, DVC, Arize, Comet ML, Qwak, Databricks, MLFlow, Apache Spark
LLM:	OpenAI, Anthropic, Azure, Llama-3, Mistral, Multi-modal LLM(TTS/STT/VST/AST), SDXL, Gemini, Vertex, Perplexity, Advanced RAG, TAG, Advanced chunking strategy
Fine-tuning&Training: Benchmarks:	Transformers, Unsloth, LitGPT, FL(with ONNX Runtime) lighteval, evals, ragas, Perforator
Data Science:	Pydantic, PySpark, Pandas, Polar, Ibis, BigQuery
WebFramework:	Django, Flask, FastAPI, Express.JS, Node.js, PHP, React.JS, Next.JS
Database:	PostgreSQL, MongoDB, Aurora DB, DynamoDB, Redis, Qdrant, Snowflake, Hopworks, PGVector, Pinecone, Milvus
Orchestrator:	Docker Swarm, ECS, K8s, Airflow, Kubeflow, ZenML, PipeDream
ORM:	Alchemy, Peewee, Django ORM
API Design Models:	REST API, RPC, GraphQL
CI/CD:	Git, GitLab, GitHub Actions, Jenkins, Kubernetes, CircleCI
Cloud Infrastructure Tools:	Terraform, Cloudformation, CDK, Pulumi
Streaming:	Apache Kafka, AWS Flink, Bytewax, CDC pattern, RabbitMQ, GCP pub/sub
ML Optimization	TorchServe, TensorFlow Serving, Ray Serve, NVIDIA TensorRT-LLM, NVIDIA Triton Inference Server, ollama, llama.cpp, vllm, sglang, LitServe, TGI, KV cache Continuous batching, Speculative Decoding, FL(with ONNX)

- Career Experience

InsoftAI, FL, United State

Senior Machine Learning & AI-Agent Engineer & Backend Engineer

02/2023 – present

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. Developed Support-nGen™, a proprietary system designed to enhance customer service by efficiently managing FAQs, support tickets, and complex queries.

Also, I developed a LLM Twin, an advanced AI character that emulates individual writing/coding styles, voices, and

personalities to serve as an effective writing co-pilot, facilitating brand creation by automating the writing process, generating new creative ideas, and streamlining content creation.

- ☑ Implemented expertise in implementing a sequential request processing system with a strong emphasis on low latency, adopting an online real-time inference deployment architecture to enhance overall performance and responsiveness.
- ☑ Designed cloud-service/microservice architecture by splitting the ML service into a REST API server for business logic and an optimized LLM microservice, leveraging powerful machines and various engines to enhance latency and memory usage, thereby facilitating quick adaptation of the infrastructure based on different LLM sizes.
- ☑ Demonstrated a comprehensive approach by integrating Graph RAG with Neo4j within the business microservice, incorporating advanced RAG techniques to optimize the pre-retrieval, retrieval, and post-retrieval steps, resulting in enhanced accuracy and improved response, implementing binay quantization solution improving RAG search to 40x faster.
- ☑ Utilized Opik to develop a sophisticated dashboard for monitoring complex prompt traces and implemented its Python SDK to effectively evaluate agentic and RAG applications, resulting in enhanced experiment tracking and improved performance comparisons.
- ☑ Implemented a highly efficient deployment strategy for the LLM microservice on AWS SageMaker, utilizing Hugging Face's Deep Learning Containers (DLCs) to enhance model inference. This robust infrastructure supported scalable, secure, and efficient real-time predictions through critical components such as SageMaker endpoints, model configuration, and inference components. By leveraging the Text Generation Inference (TGI) engine, the system achieved superior computational efficiency via tensor and dynamic batching for leading open-source LLMs like Mistral, Llama, and Falcon, accomplished optimizing performance with flash-attention, minimizing model size through model parallelism and weight quantization, enhancing throughput with speculative decoding, continuous batching, accelerating weight loading using safetensors, and enabling real-time interactions via token streaming, culminating in a responsive and effective LLM serving solution and achieving speedups of 2-4x or more.
- ☑ Developed and implemented fine-tuning process with Unsloth, a high-performance library, utilizing custom kernels, accelerating training by 2-5x and significantly reducing memory usage by up to 80%.
- ☑ Engineered a business microservice using FastAPI, initially deployed to AWS Elastic Kubernetes Service (EKS) or AWS Elastic Container Service (ECS), involving Dockerization of the application, pushing the Docker image to AWS ECR, and configuring the deployment, while also orchestrating ML pipelines using ZenML / Airflow, storing and versioning ML pipelines as outputs, and attaching metadata to artifacts for better observability.
- ☑ Utilized advanced profiling tools to identify costly lines of code and uncover performance blind spots in local programs and Kubernetes clusters running on Linux, successfully optimizing CPU, GPU, and I/O performance, which led to an estimated 20% reduction in infrastructure costs.
- ☑ By integrating Ragas's strengths in production monitoring and LLM-assisted metrics with ARES's configurable evaluation process and classifier-based assessments, enhanced evaluation capabilities, achieving quick iterations and in-depth, customized evaluations that significantly improve performance outcomes.
- ☑ Designed a clean backend–frontend architecture with FastAPI, built a web API using FastAPI and add WebSocket support agent can respond in real time.
- ☑ Exhibited strong leadership abilities by mentoring junior staff, enhancing their communication skills, and encouraging professional development.

Brainhub, Gliwice, Poland

AI/MLOps Engineer & Tech Lead

10/2019 – 12/2022

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

I engineered a real-time personalized recommender system for H&M fashion articles using the 4-stage recommender architecture and a two-tower model design architecture, leveraging the Hopworks AI Lakehouse.

Also, implemented an Intelligent Financial Advisor that provides real-time insights and personalized recommendations to users based on their financial goals and market conditions. Built an ML system for forecasting hourly energy consumption levels across Denmark and built healthcare insurance fraud identification using PCA anomaly detection using unsupervised anomalous outlier techniques on a minimal set of metrics made available in the CMS Medicare inpatient claims.

- ☑ Designed 4-stage architecture to build a system that can handle recommendations from a catalog of millions of items and two-tower model, a flexible neural network design that creates embeddings for users and items and optimized deploying ML models using Auto scaling, model optimization/parallelism/quantization, implementing a strategy similar to what TikTok employs for short videos, which will be applied to H&M retail items.
- ☑ Enhanced recommender systems by integrating advanced evaluation metrics such as NDCG, Precision@K, Recall@K, and Mean Reciprocal Rank (MRR), providing nuanced insights into model performance and user relevance, ultimately improving user satisfaction and engagement.
- ☑ Deployed real-time recommendations using Hopsworks Serverless and KServe, a runtime engine for serving predictive and generative ML models on Kubernetes, which simplifies autoscaling, networking, health checks, and server configuration while providing advanced features like GPU autoscaling and canary rollouts; through KServe, I successfully implemented two distinct services— the query encoder service and the ranking service—resulting in improved model performance and responsiveness in production.
- ☑ A real-time streaming pipeline that listens to financial news, cleans & embeds the documents, and loads them to a vector DB.
- ☑ A fine-tuning pipeline (deployed as a serverless continuous training) that fine-tunes an LLM on financial data using QLoRA, monitors the experiments using an experiment tracker, and saves the best model to a model registry.
- ☑ An inference pipeline built in LangChain, LangGraph(multi-agents) and LangSmith (deployed as a serverless RESTful API and monitoring) that loads the fine-tuned LLM from the model registry and answers financial questions using RAG/TAG (using the vector DB populated with financial news in real-time) .

Kensho, Massachusetts, United State

Backend-heavy AI Developer

09/2016 – 09/2019

I worked on a TTS and STT solution, exposing it as an API that accurately clones voices from a short audio clip, significantly enhancing user experience in voice synthesis applications, and built an ML system for forecasting hourly energy consumption levels across Denmark, improving predictive accuracy and operational planning.

- ☑ I built an inference pipeline in LangChain as a serverless RESTful API, enabling real-time financial question answering using RAG/TAG, significantly improving user engagement.
- ☑ Extended Meta's Llama 3 model with multimodal projector, allowing direct audio input for faster responses compared to traditional ASR-LLM combinations, enhancing system efficiency.
- ☑ Implemented multi-modal techniques to handle diverse input data types, enhancing the system's versatility and user experience, leveraging Neo4j for graph-based RAG, significantly improving data retrieval speeds and accuracy, automated complex workflows using Apache Airflow, resulting in a 50% reduction in processing time and increased reliability in data handling.
- ☑ Designed and launched a comprehensive API ecosystem for white-label roadside assistance platforms, enabling seamless integrations with insurance providers and automotive OEMs while improving thirdparty onboarding efficiency by 50%.
- ☑ Built efficient batch prediction pipelines using Python, leveraging a Feature Store and GCS, orchestrated with Airflow, resulting in streamlined predictions and improved operational workflows.

Dana Scott Design, Indianapolis, United States

Full Stack Developer/assistant

02/2014 – 8/2016

I converted 24 design mockups into highly-quality, pixel-perfect code using React.js, enhancing the visual consistency and user interface of applications.

- ☑ Assisted in product development, boosting customer satisfaction by 17% through upgrades to existing suites, significantly improving user retention.
- ☑ Optimized a JavaScript function for complex mathematical operations, enhancing performance and efficiency through my strong mathematical background and JavaScript proficiency.

- Education

Bachelor Degree in Computer Science

University of Kansas (2009 – 2013)