

Bobby Estes

Senior Python/Gen AI/ML Software Engineer & Agent Specialist

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

02/2023 – present

I led design and implementation of a multimodal AI healthcare multi-agent system to automate multi-specialist medical diagnosis. Expanded the system's capabilities by integrating additional medical specialties, including dermatology and nephrology, while establishing seamless connections with electronic health record (EHR) systems to enable direct report retrieval and analysis. I trained AI model using custom-trained algorithms to enhance the accuracy of medical analyses,

significantly improving diagnostic efficiency and collaboration among specialists. Also, developed a robust architecture comprising medical report extraction, specialist AI agents for multi-perspective analysis, and a multidisciplinary team agent to aggregate insights, ultimately generating comprehensive multidisciplinary diagnoses stored as text files.

- ☑ 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 a cloud-service/microservice architecture for an AI-powered healthcare system, 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, facilitating quick adaptation of the infrastructure based on different LLM sizes.
- ☑ Expanded system capabilities by incorporating multiple medical specialties, including dermatology and nephrology, while integrating with EHR systems to enable seamless report retrieval and analysis, enhancing the system's functionality and accessibility.
- ☑ Implemented a comprehensive approach using LangChain and GPT-4o to analyze medical reports from diverse specialist perspectives such as cardiology, psychology, pulmonology, neurology, endocrinology, and immunology, aggregating insights into a final multidisciplinary diagnosis that improves patient care.
- ☑ Fine-tuned LLM using custom-trained algorithms, significantly enhancing the accuracy of medical analyses and streamlining traditional diagnosis processes, allowing healthcare professionals to focus on critical cases and improving overall efficiency.
- ☑ 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.
- ☑ Built a robust architecture consisting of medical report extraction, specialist AI agents for multi-perspective analysis, and a multidisciplinary team agent that aggregates insights, ensuring comprehensive and accurate diagnoses stored as text files.
- ☑ Utilized advanced profiling tools to identify performance bottlenecks in the healthcare system, optimizing CPU, GPU, and I/O performance, which led to an estimated 20% reduction in operational costs and enhanced system responsiveness.
- ☑ Developed and implemented a highly efficient deployment strategy for the LLM microservice on AWS SageMaker, utilizing DLCs to enhance model inference, supporting scalable, secure, and efficient real-time predictions.
- ☑ Orchestrated ML pipelines using ZenML / Airflow, storing and versioning ML pipelines as outputs, and attaching metadata to artifacts for better observability and management of the healthcare system's workflows.
- ☑ Implemented robust security protocols compliant with HIPAA and GDPR, conducting comprehensive vulnerability assessments that reduced security risks by 30%. Facilitated training sessions to enhance cybersecurity awareness among team members and medical staff, promoting a proactive security culture.
- ☑ Collaborated with medical specialists, cross-functional team to gather insights, understand unique challenges, and ensure the system met clinical needs effectively.
- ☑ Designed a clean backend–frontend architecture with FastAPI, built a web API using FastAPI and add WebSocket support agent can respond in real time.

Brainhub, Gliwice, Poland

AI/MLOps Engineer & Backend & Agent Developer

10/2019 – 12/2022

I developed Sierra.ai which redefines how businesses interact with data by simplifying document management and information accessibility. Sierra's customer-centric approach establishes it as a trusted and secure partner for businesses of all sizes looking to implement multi-agent systems with langchain/langGraph.

Also, 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 Hopsworks AI Lakehouse.

- ☑ I was responsible for the development of Sierra.ai, revolutionizing document management and information accessibility for businesses, resulting in a 30% increase in operational efficiency for clients, applied extensive knowledge

in AI/ML research to design and implement advanced algorithms, enhancing the platform's capability to process and analyze complex data sets effectively.

- ☑ Led the formulation and execution of technical strategies that align with business goals, contributing to a 25% growth in user adoption rates over the past year, designed and optimized multi-AI agents capable of autonomous decision-making, which improved response times by 40% and reduced manual intervention needs.
- ☑ I designed and implemented a modular Python package that orchestrates the ML workflow into three fully automated real-time pipelines—feature, training, and inference—while reducing processing time by 62.5% preserving the accuracy.
- ☑ By addressing the limitations of traditional RAG with KAG, achieved over 94% accuracy in popular science queries and 93% in interpreting medical indicators, showed similarly impressive results, with precision rates of 91.6% and recall rates of 71.8% — a significant improvement over traditional RAG methods.
- ☑ I 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 and deployed a scalable and cost-effective asynchronous batch architecture on AWS ECS and SQS, dynamically scaling based on job volume and achieving a 52% reduction in AWS costs.
- ☑ 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.

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.
- ☑ 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)