Bobby Estes

Senior AI Research Manager

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

I am a dedicated AI Research Manager with over a decade of experience in software development and AI innovation. My career has been driven by a passion for advancing AI technologies, particularly in generative AI and AI-driven financial solutions. I excel in leading research projects from inception to deployment, ensuring alignment with business objectives and fostering a culture of innovation and continuous learning. My expertise bridges the gap between cutting-edge AI research and real-world applications, delivering transformative results.

SKILLS

Languages: Python, JavaScript, 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

Data Science: Pydantic, PySpark, Pandas, Polar, Ibis, BigQuery

Web Framework: Django, Flask, FastAPI, Express.JS, Node.js, PHP, React.JS, Next.JS

Database: PostgreSQL, MongoDB, Aurora DB, DynamoDB, Redis, Qdrant, SnowFlake, Hopsworks, 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

Other: Continuous batching, Speculative Decoding, FL (with ONNX)

PROFESSIONAL EXPERIENCE

Clearwater Analytics - AI Research Manager

02/2023 - Present

- Lead the design and implementation of GenAI models and algorithms, enhancing financial and investment management solutions, resulting in a 30% increase in processing efficiency.
- Collaborate with cross-functional teams to integrate AI research findings into product development, driving a 25% improvement in product innovation and market competitiveness.
- Oversee research projects from ideation to deployment, ensuring timely delivery and alignment with strategic goals, achieving a 20% reduction in project delivery time.
- Build and maintain partnerships with academic institutions and industry leaders, fostering research collaborations that have led to a 15% increase in knowledge sharing and innovation.
- Implement strategic research roadmaps aligned with company objectives, contributing to a 40% growth in AI-driven product offerings.
- Mentor and develop a high-performing team of AI researchers and engineers, enhancing team productivity by 35% through effective leadership and skill development initiatives.

InsoftAI - Senior Machine Learning Engineer

02/2023 - Present

- Spearheaded the development of AI-driven platforms, streamlining business operations by automating 90% of customer inquiries, resulting in a 50% increase in customer satisfaction.
- Engineered a proprietary AI system, Support-nGen, to optimize customer service processes, reducing response times by 40% and enhancing service efficiency.
- Designed and deployed cloud-based microservices architecture, improving system scalability and reducing latency by 30%.
- Integrated advanced RAG techniques with Neo4j, achieving a 40x improvement in search speed and accuracy.
- Optimized LLM microservice deployment on AWS SageMaker, enhancing model inference efficiency by 200% through dynamic batching and model parallelism.

- Mentored junior engineers, fostering a collaborative environment that improved team communication and project outcomes by 25%.

Brainhub - AI/MLOps Engineer

10/2019 - 12/2022

- Developed Sierra.ai, revolutionizing document management and increasing client operational efficiency by 30%.
- Led technical strategy formulation, aligning with business goals and driving a 25% growth in user adoption.
- Designed multi-AI agent systems, reducing response times by 40% and enhancing decision-making capabilities.
- Implemented scalable ML architectures, achieving a 62.5% reduction in processing time while maintaining accuracy.
- Adopted MLOps best practices, reducing AWS costs by 52% through efficient resource management.
- Enhanced recommender systems, improving user satisfaction by integrating advanced evaluation metrics.

Kensho - Backend-heavy AI Developer

09/2016 - 09/2019

- Developed TTS and STT solutions, improving voice synthesis accuracy and user experience.
- Built a real-time financial question-answering API, enhancing user engagement by 50%.
- Extended Llama 3 with multimodal capabilities, reducing response times and increasing system efficiency.
- Designed streaming pipelines for financial data, improving retrieval efficiency by 30%.
- Implemented a continuous training solution for financial LLMs, optimizing model performance and adaptability.

Dana Scott Design - Full Stack Developer/Assistant

02/2014 - 08/2016

- Transitioned from front-end to full-stack development, enhancing application functionality and user experience.
- Developed reusable React components, improving productivity and application scalability.
- Designed RESTful APIs, optimizing data flow and communication for enhanced application performance.

EDUCATION

University of Kansas - Bachelor's Degree in Computer Science

2009 - 2013