AUSTIN LALLY

Agentic Systems · Retrieval · Generative Pipelines

AI engineer specializing in production LLM systems, multi-agent orchestration, and large-scale search and ranking systems. Built and deployed AI systems across multimodal retrieval, real-time conversational agents, and generative pipelines, with a focus on evaluation-driven optimization. Director and founder experience.

AI Engineer · Cortina Productions · McLean, VA · 2024 – Present

Agentic Conversational AI — Theodore Roosevelt Presidential Library

- Architected a production multi-agent LLM system for a permanent installation, with a primary TR-voice agent and supporting agents for classification, context tracking, and note-taking.
- Engineered real-time LLM orchestration system (WebSockets) handling kiosk client connections to the agent pipeline, prompt assembly, context management, and low-latency streaming responses via Azure Speech.
- Implemented a layered guardrail system including upstream input classification, downstream output inspection, and a fixed top-level prompt layer enforcing safety constraints across CMS-driven content.

Intelligent Search & ML Ranking — National Archives

- Architected a semantic retrieval and ranking system powering 30 interactive gallery stations over a corpus of 15M archival records.
- Trained engagement-based ranking model to filter corpus to 2M high-quality records and provide query-time ranking signals.
- Engineered multimodal ingestion pipeline with LLM-based enrichment and CLIP/BGE embeddings to build retrieval indexes.
- Built offline filtering and query-time ranking system using user-labeled feedback collected via a purpose-built evaluation interface.

Production AI Engineering

- Built production APIs for identity-conditioned image generation across generative model variants with consistent likeness quality, deployed via Terraform on AWS (EB, S3, CloudFront).
- Developed SFace-based evaluation system to score identity similarity, enabling model evaluation and runtime quality gating with retry.
- Built LLM-driven pipeline tagging transcripts per-utterance and matching to editorial briefs to surface candidate sound bites for producer review.
- Trained BERT/RoBERTa classifiers for profanity detection as a content safety guardrail; optimized for inference via ONNX on .NET 8.

Founder · WxH Inc. · Corvallis, OR · 2021 – 2024

- Built a vision-based iPad app for art arrangement visualization.
- Trained room geometry prediction models from current research papers.

Director of Engineering · Concentric Sky · Eugene, OR · 2011 – 2019

- Progressed from engineer to Director; defined technical architectures on AWS (EC2/ECS, RDS, S3, CloudFront) for web-facing client products, led distributed cross-functional teams, anchored technical decision-making.

Skills

AI/ML
PyTorch · HuggingFace Transformers · ONNX · BERT/RoBERTa · CLIP

LLM & Generative AI
Multi-agent orchestration · LLM guardrails · AutoGen · RAG · Diffusion models · ComfyUI · Evaluation & quality gating

Search & Retrieval
Semantic / hybrid search · Embedding pipelines · Azure AI Search · FAISS · Reranking models

Backend & Infra
Python · FastAPI · Docker · Azure · AWS · Databricks · Streaming & async APIs

Frontend
Real-time streaming · TypeScript · WebSockets

Education

MS, Computer Science
Machine Learning / AI
Oregon State Univ. · 2022

BS, Computer Science
Univ. of Oregon · 2011

Contact

 austin-lally

 austinlally.com

 austin@austinlally.com

 (541) 514-1015

 Washington DC area