

Rajasekhar Mekala

Santa Clara, CA · +1 (949) 992-6595 · rajasekharreddymekala732@gmail.com · rajasekharmekala.github.io · linkedin.com/in/rajasekhar-mekala · github.com/rajasekharmekala

Professional Summary

Senior Software & Machine Learning Systems Engineer with 5+ years of experience architecting enterprise Generative AI platforms, multi-tenant LLM agent runtimes, and high-throughput retrieval infrastructure. Core contributor to Adobe Brand Concierge, designing end-to-end agent orchestration (PydanticAI, MCP), neural source-attribution engines, ONNX-optimized inference microservices, and automated prompt-optimization pipelines. Published NLP researcher (NAACL, EMNLP) with a Master's in Computer Science from UC Irvine (GPA 3.94) and B.Tech from IIT Kharagpur (AIR 481).

Work Experience

Adobe

Software Engineer — Generative AI Platforms & Agent Systems

Santa Clara, CA

Jun 2023 – Present

- **Agent Orchestration Engine (BCOS):** Architected the next-generation Brand Concierge Operating System, unifying multi-hop legacy pipelines into a single high-performance **PydanticAI** agent runtime with declarative manifest compilation, supporting multi-tenant enterprise workflows via Agent-to-Agent (A2A) JSON-RPC protocols.
- **Declarative Tool & Capability Infrastructure:** Designed an extensible Pydantic tool configuration subsystem, introducing declarative pre/post execution hooks, state capture primitives, and configurable **ToolErrorPolicy** for deterministic recovery under tool failures.
- **Neural Source Attribution Engine:** Implemented a factory-based attribution and citation system integrating remote cross-encoder rerankers (**bge-reranker-v2-m3**, **e5-curriculum**), client-side normalization, token truncation safeguards (<512 tokens), and dynamic request batching, delivering grounded factual citation across pilot deployments (NVIDIA, Synopsys, ITC, DSG).
- **High-Performance Embedding & Reranking Microservices (Lucid Lens):** Built CPU-only Text Embeddings Inference (TEI) microservices on **ONNX Runtime (ORT)** with dynamic axes, overcoming the 4-batch Candle limitation to achieve full-batch reranking of 30+ candidate passages with sub-second p50 latency (588ms–796ms).
- **LLM Manifest Routing & Model Profiles:** Built a dynamic named LLM profile system enabling zero-downtime, manifest-driven runtime model routing across Azure OpenAI (GPT-4.1 / GPT-4o), OpenAI, and private enterprise model endpoints.
- **Agent Optimization & Eval Studio (BC Agent Optimizer):** Engineered high-throughput evaluation and prompt compilation backend integrating **GEPA** (Genetic-Pareto Prompt Optimization) and **DSPy** for reflective manifest compilation, achieving Pareto-optimal prompt mutations with 35x fewer rollouts than RL baselines.
- **Evaluation Dataset Persistence Subsystem:** Developed high-throughput MSSQL-backed dataset infrastructure supporting chunked persistence, bulk row staging, real-time build progress tracking, and span-aware execution trace evaluators.
- **Streaming Runtime & Traffic Control (BC Runtime):** Hardened the core FastAPI/MongoDB conversation service, implementing real-time SSE token streaming, UI artifact state replacement (**StreamEventProcessor**), and multi-tenant traffic isolation (**TrafficControlEnforcer**).
- **Semantic Routing Cache (BC Cache):** Engineered a standalone Q&A routing cache using local **fastembed** (**bge-small**) and semantic clustering over historical interaction logs, serving recurring brand queries with sub-millisecond lookups and reducing LLM inference costs.

TruckX Inc

Software Engineer

Sunnyvale, CA

Jul 2022 – Jun 2023

- Engineered resilient event-driven microservices processing high-frequency real-time IoT vehicle telemetry data at scale.
- Devised end-to-end telemetry solutions to track remote driving behaviors and enforce regulatory Hours-of-Service (HOS) compliance.
- Led the core engineering team in the successful architecture and production launch of ELD (Electronic Logging Device) trackers for the Canadian market.

- Investigated in-context learning robustness, prompt sensitivity, and chain-of-thought reasoning consistency in Large Language Models.
- Developed **EchoPrompt**, instructing models to rephrase queries prior to reasoning, substantially improving few-shot arithmetic and logical reasoning performance across LLMs (Published at NAACL / ARR 2023).
- Co-developed **Snoopy**, an interactive diagnostic framework exploring the causal impact of pretraining term frequencies on few-shot language model outputs (Published at EMNLP 2022).
- Served as Teaching Assistant for Senior Software Design (15 industry-sponsored projects, 75+ students).

- Spearheaded backend engineering for Adobe Documentation Help Center on Adobe Experience Manager (AEM), creating microservices for enterprise client documentation.
- Designed high-throughput search autocomplete and search analytics for Adobe RoboHelp Server, reducing search latency by **27%** and improving relevancy by **31%**.
- Received the Adobe Spot Award and Special Contribution Award for key algorithmic and architectural contributions.

- Built an ML framework on Apache Spark/Hadoop (AWS) to detect and mitigate multi-tenant noisy-neighbor performance degradation in shared cloud infrastructure; won **Best Intern Project Award** (Ranked 1st among 23 projects) and co-authored **US Patent US20190354388**.

Publications & Patents

- **NAACL / ARR 2023:** *EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning*. Rajasekhar Reddy Mekala, Yasaman Razeghi, Sameer Singh.
- **EMNLP 2022:** *Snoopy: An Online Interface for Exploring Pretraining Term Frequencies on Few-Shot LM Performance*. Yasaman Razeghi*, Rajasekhar Mekala*, Robert Logan, Matt Gardner, Sameer Singh.
- **US Patent US20190354388:** *Tenant-side detection, classification, and mitigation of noisy-neighbor induced performance degradation*. S. Mitra, S. Khosla, S. Mehta, M. R. S. Reddy, A. D. Shah.
- **Invention Disclosures (Adobe Brand Concierge):** Demand-paged agent context with revocable tool capabilities; Evidence-linked causal evolution of typed agent manifests; Anti-collusion dual optimization for judges and agents.

Education

University of California, Irvine — Master of Science in Computer ScienceIrvine, CA · Sep 2021 – Jun 2023
(GPA: 3.94 / 4.00)
Coursework: Deep Generative Models, Statistical NLP, Machine Learning, Distributed Systems, Middleware, Operating Systems, Networks

Indian Institute of Technology, Kharagpur — B.Tech in EE, Minor in Computer ScienceKharagpur, India · Jul 2014 – May 2018
(GPA: 9.04 / 10.00)
Honors: All India Rank 481 in IIT-JEE Advanced (Top 0.03% among ≈1.5M applicants); Selected for International Maths Olympiad Training Camp (2013)

Technical Skills

AI & Agent Run-times:	LLM Agent Architectures, PydanticAI, Model Context Protocol (MCP), ReAct, A2A Protocol, DSPy, GEPA, Tool Calling, Multi-Agent Orchestration, Prompt Optimization, Agent Sandboxing.
Retrieval & Inference:	Grounded RAG, Cross-Encoder Reranking (bge-reranker, e5-curriculum), ONNX Runtime (ORT), Text Embeddings Inference (TEI), FastEmbed, Source Attribution, Vector Indices, Semantic Caching.
Evaluation & ML:	LLM-as-a-Judge, Prediction-Powered Inference (PPI), Process Reward Models (PRMs), Span-level Evaluators, PyTorch, HuggingFace Transformers, Scikit-Learn.
Backend & Systems:	Python (FastAPI, asyncio, Pydantic), Java, C++, TypeScript, Server-Sent Events (SSE), REST, JSON-RPC, Kafka, Docker, Kubernetes, Linux/Unix, OpenTelemetry.
Data & Storage:	Microsoft SQL Server (pymssql, bulk staging), MongoDB, PostgreSQL, Redis, Apache Spark.