

Manoj Tummala

213-263-8957 | manojtummala0120@gmail.com | LinkedIn - [manoj-tummala](#) | Github - [manojtummala](#) | [manojtummala.github.io](#)

TECHNICAL SKILLS

Programming: Python, C/C++, SQL (advanced – window functions, CTEs), Java/Scala, Swift, JavaScript, TypeScript, PHP, C#
Distributed Systems & Networking: Kafka, Microservices, distributed data pipelines, gRPC
Data Engineering & Storage: Apache Spark, Airflow, Hadoop, ETL/ELT Pipelines, Delta Lake, Snowflake, BigQuery, MongoDB, DynamoDB
Cloud & Infra: AWS (S3, Lambda, EMR), GCP (BigQuery), Docker, Jenkins, CI/CD
Monitoring & Reliability: Log tracing, Alerting, Retry mechanisms, Fault tolerance, Operational resilience
AI & Retrieval Systems: RAG, semantic chunking, embedding-based search, LLM-as-judge evaluation.
Tools & Frameworks: Docker, Git, Flask/FastAPI(internal data services), Vector Databases (Qdrant), LLM APIs, gdb.

EXPERIENCE

Research Assistant

University of Southern California - Modular Agent-Based Task Executor for Knowledge Workflows June 2025 - Present

- Investigating modular decomposition strategies for knowledge-intensive tasks by designing and evaluating a hierarchical multi-agent architecture, exploring tradeoffs between agent specialization and **inter-agent coordination overhead**.
- Evaluating **LLM-as-judge** scoring and Chain-of-Thought prompting as mechanisms for **self-correction** in agentic loops, with preliminary findings showing improved answer consistency on multi-hop document reasoning benchmarks.
- Prototyping a **fault-tolerant** agent orchestration framework with pluggable tool interfaces (retrieval, search, structured query), informing design decisions around reproducibility and reliability in knowledge workflow systems.

Full-Stack & Infra Developer

ReviewsVerse May 2024 - May 2025

- Designed and deployed a scalable backend data-serving platform supporting 1,000+ concurrent users while enforcing strict SLIs/SLOs (<200ms latency) and system availability and emphasizing reliability and operational stability.
- Built Kafka-S3 **ingestion pipelines** with **OpenTelemetry tracing**, retries, and circuit breakers, eliminating failure propagation and boosting pipeline uptime by 30%.
- Reduced ETL latency** by 20% through **query bottleneck analysis** and engineered **metric anomaly detection** checks, eliminating false alerts by 30% to improve operational signal-to-noise ratio

Data Engineer & React JS Developer

SchoolTalkz May 2022 - May 2023

- Automated **Airflow DAGs** with retries and **real-time alerting**, eliminating manual operational toil and establish **self-healing capabilities** against transient batch failures.
- Reduced analytics reporting time by 40% by optimizing complex SQL aggregations using window functions and partition pruning, significantly **improving dashboard SLAs** for stakeholders.
- Instrumented **ETL workflows** with end-to-end telemetry and **metric collection**, to ensure **zero-loss data reconciliation** for downstream **business reporting** and accelerated incident response.

EDUCATION

University of Southern California

Masters of Science in Computer Science Aug 2023 – Aug 2025

PROJECTS

Multi-Agent Algorithmic Trading Platform

- Engineered a multi-agent system with **plug-and-play connectors** for **intraday/interday market**, news, and sentiment data, using local caching to schedule **automated strategy backtests**.
- Designed custom interfaces and an interactive **dashboard to organize multi-asset product baskets**, enabling **portfolio hedging** across complementary financial instruments.
- Implemented extensible **brokerage and algorithm interfaces** featuring strict **human-in-the-loop approval** controls prior to executing buy/sell order transactions.

On-Device Agentic Productivity OS Assistant

- Built a fully local, privacy-preserving agentic desktop assistant using Phi-3 Mini, MLX, and a custom RAG pipeline, enabling multimodal ingestion, semantic search, and tool-calling automation through a Rust-based Tauri backend and React UI.
- Optimized local LLM execution and vector search indexing to maintain sub-second response times, implementing graceful fallback handling for system resource constraints.

Self-Improving Retrieval-Augmented Generation Pipeline

- Built a retrieval-augmented generation pipeline with automated evaluation using LLM-as-judge scoring, enabling iterative tuning of chunking and retrieval parameters to improve answer consistency and factual grounding across document-based queries.

Open Data Analytics Platform

- Built a production-style analytics pipeline over public datasets, handling schema evolution, late-arriving data, and metric consistency through layered SQL models and data quality checks.