

GHANASHYAM VINOD TATTI

San Jose, CA, USA | ghanashyamvtatti@gmail.com | linkedin.com/in/ghanashyam-vt | github.com/ghanashyamvtatti | gtbuilds.dev

Senior Software Engineer — Agentic AI Platforms, LLM Evaluation & Distributed Systems | Python | MCP | Kubernetes

PROFESSIONAL SUMMARY

Engineer with 8+ years of production systems and platform-level ownership of enterprise agentic AI. Architected the **Model Context Protocol** tool layer and agentic pipelines behind Cisco Catalyst Center's AI Assistant (80% engagement lift), contributed runbooks, skills and tools to its Deep Reasoning harness, and played a vital role in the 200+ enterprise-customer beta. Owns the evaluation practice that gates every release — golden datasets, LLM-as-judge, regression gating — used it to single-handedly migrate production from GPT-4.1 to GPT-5.4-mini, and cut agent latency 40% and token spend 60%. Multiplies teams through mentoring and an org-wide prompt library; shipped an on-device LLM assistant to both app stores solo.

TECHNICAL SKILLS

Agentic AI: Model Context Protocol (MCP), agent harnesses, runbooks and skills, tool-contract design, risk-tiered action validation, LangGraph, LangSmith, RAG, deep-reasoning agents

Evaluation & Release: Golden datasets, LLM-as-judge, trajectory evaluation, regression gating, model-migration evals, prompt and tool-definition iteration, exact-revision release promotion

Systems & Performance: Python (FastAPI, asyncio, concurrency), Java, Go, Kafka, caching, rate limiting, SSE streaming, latency and token-cost profiling, distributed tracing

Infrastructure: Docker, Kubernetes, AWS, GCP, Azure, Postgres, MySQL, Elasticsearch, Grafana

PROFESSIONAL EXPERIENCE

CISCO SYSTEMS — Software Engineer, Catalyst Center

San Jose, CA | Jun 2021 – Present

- Architected core Model Context Protocol (MCP) tools and agentic pipelines for the AI Assistant platform, expanding system capabilities to drive a 80% increase in user engagement.
- Contributed to the agentic harness powering Catalyst Center's Deep Reasoning mode, authoring runbooks, agent skills, and tools that expanded its autonomous network troubleshooting and root-cause coverage.
- Led the migration of the AI Assistant from GPT-4.1 to GPT-5.4-mini end to end: designed and ran the evaluation suite, then rewrote prompts, tool definitions, and agent instructions to hold answer quality while reducing inference cost.
- Played a vital role in the Deep Reasoning beta program, now spanning 200+ enterprise customers — provisioned access, ran structured feedback cycles, and converted findings into prioritized product changes that I designed and shipped.
- Optimized deep reasoning agent infrastructure by resolving critical bottlenecks, reducing overall latency by 40% and token consumption by 60%.
- Engineered a centralized, organization-wide AI prompt library and led training workshops, driving a 45% increase in tool adoption and saving teams an average of 5 hours per week.

CISCO SYSTEMS — Software Engineer Intern, Assurance

San Jose, CA | Jun – Aug 2020

- Pioneered an Anomaly Detection system for automated Root Cause Analysis of customer bugs, reducing process duration by over 80%.

THOUGHTCLAN TECHNOLOGIES — Machine Learning Engineer

Bangalore, India | Sep 2016 – Aug 2019

- Engineered an Automated Machine Learning Solution for Industrial IoT, achieving >90% accuracy and enabling 10-second remote deployment of efficient TensorFlow Lite models across diverse data types and mobile/edge platforms.
- Pioneered a Low-Cost Gas Leak Detection System by training a Convolutional Neural Network on infrared images, achieving 85% detection accuracy and reducing client costs by 98%.
- Trained new recruits quarterly in advanced Python, Machine Learning fundamentals, and Git, accelerating technical onboarding and enhancing team proficiency.

RARE MILE TECHNOLOGIES — Software Engineer

Bangalore, India | Jul 2015 – Sep 2016

- Engineered social media integrations, a platform-wide search with autocomplete, and an XLSX import tool for Social Survey, onboarding hundreds of thousands of new customers in seconds.

PROJECTS

- **THINKALOUD** (ghanashyamvtatti.github.io/thinkaloud-site) — Fully on-device, privacy-first AI assistant (macOS, Windows, iOS, Android): local quantized LLMs via llama.cpp and LiteRT-LM, zero telemetry, a reasoning-trace UI that surfaces the model's intermediate thinking, on-device semantic search, and full model-licensing review plus store release ownership.

EDUCATION

New York University — Master of Science, Computer Science (GPA 3.67)

2019 – 2021

Visvesvaraya Technological University — Bachelor of Engineering, Computer Science and Technology (GPA 3.39)

2011 – 2015