

NEERAJ BHARGAV RONDLA

Applied AI / Forward Deployed Engineer — Production LLM Systems · Agents & MCP · Evaluation

Jersey City, NJ · (551) 255-0106 · rondlanbr@gmail.com · linkedin.com/in/neerajbhargav · github.com/neerajbhargav · neerajbhargav.com

SUMMARY

Applied AI engineer who ships production LLM systems inside someone else's business and owns the outcome. Sole engineer on five live services at a NYC brokerage: I run discovery with a non-technical principal, turn vague asks into signed numeric acceptance criteria, and report results in plain English. Deep Claude API practice, published MCP servers for Claude Code and Cursor, and an evaluation discipline that has twice made me kill my own plausible improvements because the measurement said no.

EXPERIENCE

Applied AI Engineer, Landair Advisors — New York, NY

Jun 2026 – Present

- **Discovery and requirements** — converted an ambiguous “make our owner data better” ask into four signed KPIs with numeric thresholds and labeled validation sets; renegotiated the success metric with a non-technical principal after measurement showed two defensible definitions differed by 45 points, and publish plain-English status he reads without an engineer present.
- **Production LLM pipeline** — designed a 7-layer entity-resolution system (Claude Vision over scanned legal documents, recursive corporate-chain unwinding with cycle detection, permit and identity APIs) resolving any property to a named human owner; raised accuracy from a ~50% manual baseline to 93.3% over a 1,121-record run.
- **Evaluation and safety** — built labeled validation sets, longitudinal accuracy tracking in Postgres, and confidence bands derived from measured per-source accuracy rather than hand-written heuristics; low-confidence results are flagged for human review instead of returned silently, holding confidently-wrong output at zero in production.
- **Measurement discipline** — instituted a 3-run protocol after single-run variance produced two false wins, then used it to reject a larger-model upgrade that measured 14 points worse; root-caused a name-parser inversion silently corrupting 69% of records and shipped a fix that tripled exact-match accuracy (16% → 48%).
- **Delivery and adoption** — shipped the tool as a self-serve web app so brokers run their own batches, replaced manual SSH deploys with a guarded one-click workflow with health checks and rollback, and tuned batch cost to ~\$8–15 per 2,000 records.

AI Software Engineer, Innovcentric LLC — Remote

Nov 2025 – Jun 2026

- **Customer-facing delivery** — ran discovery with product owners and operations managers, decomposed ambiguous requirements into production AI architectures, validated in pilot before rollout, and codified repeatable deployment patterns; delivered white-glove onboarding with L1–L2 support, office hours, and runbooks.
- **Agents** — deployed an autonomous ReAct agent on GPT-4 function calling averaging 5–10 tool calls per query, with self-correction loops, sub-agents for specialized tasks, dual memory (conversation buffer + persistent vector store), and human-in-the-loop escalation on sensitive paths.
- **Retrieval** — production RAG pipeline (LangChain, ChromaDB, FastAPI) over 10,000+ pages; benchmarked 4 chunking strategies and 3 embedding models to reach sub-2s p50 retrieval for 100+ concurrent agents.
- **Evaluation** — RAGAS + LangSmith harness scoring faithfulness, relevance, and hallucination on every response, with drift monitoring that held quality flat through a GPT-4 → GPT-4o migration; cut factual errors ~60%, across 50+ version-controlled prompt templates carrying guardrails and injection defense.

Software Engineer, Wipro Ltd. — Hyderabad, India

Aug 2021 – Jul 2023

- **ML in production** — NLP ticket-routing pipeline (scikit-learn + HuggingFace BERT) at 91% accuracy over 5,000+ tickets/month, cutting manual triage 40%, served via containerized FastAPI.
- **Developer tooling** — AI code-review assistant (Python + OpenAI Codex) adopted by 30+ engineers that cut PR merge cycles 25%; maintained enterprise React applications serving 10,000+ users.

OPEN SOURCE — MCP & AGENT TOOLING

- **VIBETTER** — MCP server (Python, FastMCP) for Claude Code, Cursor, and Windsurf: git-diff explanation with contextual reasoning, source-grounded Q&A with file:line citations, codebase-aware debugging, interactive dependency graph. MIT. github.com/neerajbhargav/vibetter
- **OmniContext** — MCP bridge and session-handoff tool (Node.js, MCP SDK, TypeScript): a local daemon compresses git state and active tasks into handoff prompts, plus a native MCP server exposing a context://current resource to Cursor and Claude Code. Handshake verified over stdio (initialize, tools/list, resources/list). npx omni-context
- **Context Refinery** — LangGraph multi-agent context engine: intent analysis → hybrid retrieval (dense + BM25 + RRF) → cross-encoder reranking → eval-gated self-refinement scored by custom local metrics (n-gram grounding, information density, budget utilization), with optional DeepEval/RAGAS. Runs offline via Ollama. MIT. github.com/neerajbhargav/context-refinery

TECHNICAL SKILLS

LLMs, Agents & MCP: Claude API (Sonnet, Opus, Haiku, Vision), GPT-4/4o, Gemini, Llama; tool use, structured outputs, streaming, sub-agents, multi-agent orchestration, prompt engineering and versioning, prompt caching, token/cost budgeting, rate-limit engineering; Model Context Protocol, FastMCP, MCP SDK, LangGraph, LangChain, DSPy, ReAct, human-in-the-loop

Retrieval & Evaluation: RAG, agentic RAG, hybrid retrieval (BM25 + dense + RRF), cross-encoder reranking, chunking, embeddings, pgvector, ChromaDB, Pinecone, FAISS; RAGAS, DeepEval, LangSmith, LLM-as-judge, faithfulness and hallucination scoring, drift monitoring, guardrails, prompt-injection defense

Engineering: Python, TypeScript, SQL, Node.js, FastAPI, Pydantic, PostgreSQL, Redis, asyncio, Docker, AWS, GitHub Actions, Linux, React

Customer Engagement: Technical discovery, requirements decomposition, pilot-to-production, white-glove onboarding, runbooks, executive communication

EDUCATION

M.S. Computer Science, New Jersey Institute of Technology, Newark, NJ · Dec 2025

B.Tech. Computer Science & Engineering, Jawaharlal Nehru Technological University, Hyderabad, India · May 2022