

Neeraj Bhargav Rondla

Applied AI Engineer | Forward Deployed Engineer

LLM Agents · RAG · LLM Evaluation · MCP · Python · TypeScript · React

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

SUMMARY

Applied AI engineer who ships **measured, production LLM systems** and owns each one from customer discovery through evals to deploys. Sole AI engineer at an NYC real estate brokerage: built an agentic Claude pipeline that finds the real owner behind each property, LLC chains included (**97.1%** accuracy on 1,088 of 1,120 labeled properties), and merged **1,000+ PRs** across 4 repos.

EXPERIENCE

Applied AI Engineer · Landair Advisors

New York, NY · Jun 2026 - Present

- Owns five production AI services (owner resolution, sales classification, buyer ranking, valuations, call scoring) plus the broker web platform.
- Raised owner-resolution accuracy to **97.1%** (1,088 of 1,120 labeled properties, Oct 2026; strict top pick 95.5%) with an agentic Claude resolver: Claude Vision over scanned deeds, LLC-chain unwinding, record APIs called cheapest-first.
 - Built the eval layer behind model and scoring changes (labeled sets, confidence bands, 3-run protocol); low-confidence results go to human review. It killed my own Haiku-to-Sonnet upgrade when Sonnet scored **53% vs Haiku's 67%** (15-row set, 3 runs).
 - Worked with the firm's broker founders to turn **four signed KPIs** into acceptance tests (scoring rules, labeled sets); renegotiated phone accuracy when two fair definitions differed by 46 points (48% vs 94%, 98 properties).
 - Classified 17,990 comparable sales into nine asset types; labels match a broker-labeled key on **98.3%** (118/120 random) and 97.4% (149/153 stratified) when either of two labels counts (11 store rows keyed to my rule; rules tuned on these sets).
 - Rejected an LLM comp picker after a 37-property eval: a distance-plus-age sort matched **44.4%** of still-valid broker picks vs 25.0–28.5% for the LLM, so the simple rule shipped in an engine that values up to 64 properties a day (Sep 2026).
 - Built the broker web platform from scratch (React, Vercel, Python serverless): parallel property-data lookups feed one Claude call, and a model and effort test cut median report time from 41.5 s to **11.1 s** (6 addresses, Oct 2026).
 - Own production operations (641 pytest files, CI on every PR for the 2 largest services, guarded deploys with rollback); root-caused a **two-month silent outage** and a silent read cap that piled up 232K duplicate rows.

AI Software Engineer · Innovcentric LLC

Remote, USA · Nov 2025 - Jun 2026

- Served **100+ concurrent support agents** at **sub-2s p50** with production RAG over **10,000+ pages** (LangChain, ChromaDB, FastAPI), after benchmarking 4 chunking strategies and 3 embedding models.
- Cut factual errors **~60%** with a RAGAS + LangSmith harness; held quality flat through a model migration of 50+ versioned prompts.
- Deployed a ReAct agent on OpenAI function calling (**5–10 tool calls/query**) with sub-agents and human-in-the-loop escalation.
- Ran discovery with product owners and ops managers, piloted before rollout, and owned onboarding and runbooks.

Software Engineer · Wipro Ltd.

Hyderabad, India · Aug 2021 - Jul 2023

- Cut manual ticket triage **40%** with a BERT routing classifier at **91% accuracy** over **5,000+ tickets/month**.
- Cut PR merge cycles **25%** with an AI code-review tool adopted by **30+ engineers**; a portal chatbot cut repeat queries 40% for 10,000+ users.

PROJECTS & OPEN SOURCE

- **VIBETTER** · Open-source MCP server (Python, FastMCP) for Claude Code and Cursor that checks every file:line citation it parses: **82.7%** of 375 citations point to lines the model was shown (198 answers, 3 runs). MIT. github.com/neerajbhargav/vibetter
- **sidegig** · Open-source browser sidebar (TypeScript, Manifest V3, MCP tools): a rewritten sign-in-code detector raised recall from 53.3% to **77.3%** at 96.7% precision (330 synthetic held-out emails). GPL-3.0. github.com/neerajbhargav/sidegig
- **SPLITTER** · Shared expenses, exact to the cent (Next.js 16, Supabase, Row Level Security on all 23 tables): fewest payments in **500/500** seeded circle groups vs 368/500 for greedy; 341 tests. MIT. github.com/neerajbhargav/splitter

TECHNICAL SKILLS

LLMs & Agents: Claude API (Opus, Sonnet, Haiku, Vision), GPT-4o, Gemini, Ollama; tool calling, structured outputs, context engineering, prompt caching, guardrails; ReAct, sub-agents, human-in-the-loop, LangGraph, LangChain

RAG & Evaluation: RAG, hybrid retrieval (BM25 + dense + RRF), ChromaDB; golden datasets, LLM-as-judge, CI eval gates, bootstrap CIs, RAGAS, LangSmith

MCP & AI Dev Tools: MCP servers and clients, FastMCP, Claude Code (subagents, hooks), Cursor

Engineering: Python, TypeScript, SQL, FastAPI, Node.js, asyncio, REST, React, Next.js, Vue 3, PostgreSQL, Supabase, Redis

Cloud, DevOps & ML: Docker, Vercel, Linux, GitHub Actions CI/CD, pytest, Playwright, OpenTelemetry; HuggingFace, BERT, scikit-learn

Forward Deployed: technical discovery, scoping to signed KPIs, pilot-to-production, executive communication, onboarding, runbooks

CERTIFICATIONS & EDUCATION

Claude Certified Architect, Foundations (CCAR-F), Anthropic	In progress
Anthropic Academy course track, incl. Claude Code in Action, Anthropic	Jan 2026
Google UX Design Professional Certificate, Google	Jan 2024
M.S. Computer Science, New Jersey Institute of Technology (NJIT), Newark, NJ	Dec 2025
B.Tech. Computer Science & Engineering, Jawaharlal Nehru Technological University (JNTUH), Hyderabad, India	May 2022