

Vibhek Soni

Backend Engineer, AI Infrastructure

New York, NY Open to remote or hybrid in NYC

vibheksoni@engineer.com github.com/vibheksoni linkedin.com/in/vibheksoni vibheksoni.com

EXPERIENCE

FreeTheAI, Backend Engineer

Nov 2024 to Present

Go, Gin, PostgreSQL, Redis freetheai.org

- Run an OpenAI-compatible API gateway that has served over 210 billion tokens across 7.3 million requests for over 3,900 active accounts in more than 120 countries over the last 30 days.
- Built API keys and accounts, a wallet and credit system, usage metering, rate and concurrency limits, and routing by price or speed across over 110 live models.
- Recorded billable usage before any abuse check runs so enforcement can never undo or double-count it.
- Ran new abuse detectors in shadow mode where they only flag cases for human review until they are trusted.
- Added 119 Go test files across three services covering rate limits, access rules, abuse history, and the translation that lets each upstream provider answer through one OpenAI-compatible endpoint.

Macan Studios, Freelance Software Developer

Jun 2025 to Aug 2026

Remote

- Built Discord bots, websites with custom backends, and Windows desktop software for private clients.

OPEN SOURCE

stealth-browser-mcp, Creator and Maintainer

Aug 2025 to Present

Python, FastMCP, nodriver, Chrome DevTools Protocol github.com/vibheksoni/stealth-browser-mcp

- Maintain open-source Chrome DevTools Protocol infrastructure for AI agents as a Python MCP server that now has over 3,200 stars, over 400 forks, and merged pull requests from 10 outside contributors.
- Split 101 tools into 11 sections that can be switched off at startup so an agent can run a 21-tool core instead of picking from the full set.
- Added a cross-platform browser detection benchmark that runs weekly in GitHub Actions on Linux, Windows, and macOS in headed and headless mode and publishes the results to the repo.

UniClaudeProxy

Feb 2026

Python, FastAPI over 60 stars

- Built a FastAPI proxy that translates the Anthropic Messages API to OpenAI, Gemini, DeepSeek, and Ollama with streaming, tool calling, and hot-reload config.

VerbalCodeAI

May 2025

Python, Ollama, MCP over 50 stars

- Built a terminal tool that indexes a codebase locally with embeddings and answers questions like where a function is defined and it also runs as an MCP server.

EDUCATION

Western Governors University, B.S. Computer Science

Graduated Sep 2026

SKILLS

Languages Python, SQL, JavaScript, Go

Backend FastAPI, Flask, Gin, REST APIs, WebSockets, PostgreSQL, Redis, MySQL

AI and Agents MCP, FastMCP, OpenAI and Anthropic APIs, model routing, embeddings, Ollama

Infrastructure Linux, Docker, Git, GitHub Actions, Chrome DevTools Protocol