Abhinav Malkoochi

abhinav.malkoochi@gmail.com | https://github.com/AbhinavMalkoochi |6307310098

Education

University of Texas at Dallas– BS in Computer Science Aug 2023 – Dec 2025

Experience

Full Stack Intern, 'Sup – Remote Feb 2025 – July 2025

- Engineered a web scraping system using Python and Selenium to extract and analyze data from 30,000+ LinkedIn profiles
- Developed an cold email AI automation pipeline with Next.js that leveraged enriched data to generate personalized messages

LLM Research, UT Dallas – Richardson, TX Jan 2025 – May 2025

- Designed and executed large-scale experiments on context length scaling for LLMs
- Achieved 2x faster convergence and 50% lower compute vs. fixed 24k training while improving benchmark accuracy (+6% on AIME/AMC/MATH-500).
- Introduced an iterative curriculum for context scaling for more efficient long-context reasoning, lower clipping ratio, and efficient token utilization on small models

Full-Stack Engineer, Enky – Dallas, TX Dec 2024 – Present

- Developed Enky, a full-stack marketplace built with React, Express, Supabase, and Stripe, connecting 200+ artists and content creators to democratize music marketing.
- Built a secure escrow-based payment system, with automated payouts and refund mechanisms
- Led growth initiatives (SEO, social media marketing, etc) that drove 100% month-over-month user growth, ensuring consistent engagement and successful collaborations.

Artificial Intelligence Intern, XNode.AI – Remote Jun 2024 – Aug 2024

- Developed a Neo4j knowledge graph with vector embeddings, centralizing company data (product, GitHub, specs)
- Built an agentic chatbot with an LLM and RAG for knowledge graph interaction, enabling queries and insights.
- Boosted knowledge graph query accuracy by 75% via RAG implementation

Projects

Browser Agent Typescript, Python

- Built a TypeScript-based autonomous browser agent enabling LLMs to perform real web actions (DOM traversal, form fills, navigation), achieving <150ms action latency and completing multi-page workflows with 95% reliability.
- Architected a multi-session isolated runtime with deterministic action queues, sandboxed JS execution, and concurrency-safe state propagation, supporting 10 fully parallel browser agents

MCP Code Typescript

- Built a TypeScript system that let AI agents load MCP tools on-demand, reducing context usage by 90% compared to direct tool calls.
- Implemented code-execution workflows that filtered large datasets before reaching the model, cutting token overhead by 50–95% depending on workload.
- Created a filesystem-based tool interface for dozens of MCP servers, improving tool discovery speed

Skills

Languages: Java, Python, C/C++, SQL (Postgres), JavaScript/Typescript, HTML/CSS, R

Frameworks: React, Next.js, Node.js, Flask, FastAPI, Unity, Pytorch, Docker, Kafka, Redis, AWS