Umut (Hope) YILDIRIM

Full Stack - ML Engineer / New York City, NY

umut475@gmail.com +1 (415) 792-9337 umutyildirim.com in/umuthopeyildirim

Summary

I'm an ML and full-stack software engineer who turns research into customer-ready AI. I've architected, shipped production-scale features for B2C products, built intelligent agent frameworks, and spearheaded Model Context Protocol.
Skills Python, Electron, TypeScript, NextJS, Computer Vision, LLMOps, Synthetic Data, RAG, Model Context Protocol, Embeddings, ColPali, Reinforcement Learning

Competitions and Technical Projects

DOOM-Mistral - github.com/umuthopeyildirim/DOOM-Mistral (Nov 2023)

- Fine-tuned Mistral 7B base model to play DOOM(1993) based on ASCII representation.
- More detailed info at https://fireworks.ai/blog/playing-doom-using-llm

MarkAI - markai.umutyildirim.com (Nov 2023)

- An open-source OpenAI wrapper for a RAG-based chatbot that seamlessly integrates with your documents.
- All endpoints are serverless, with the exception of the database, ensuring ease of setup and immediate use.
- Utilized technologies like NextJS, OpenAI, Vercel, Clerk, Langchain and Supabase.

Flatiron Open Source - flatironopensource.com (Jan 2023)

- Built an open source hub for bootcamp graduates to have efficient access to course modules and lessons.
- Led front-end development (70%) and implemented end-to-end backend services.

Work Experience

Highlight

Full Stack - ML Engineer New York City, NY, USA (April 2024 - Present)

- **Spearheaded end-to-end delivery of Highlight's Model Context Protocol (MCP):** designed and launched the MCP Bundler to reduce plugin setup time by streamlining backend, frontend, and OS integrations; built a public marketplace for plugin discovery; and implemented critical service APIs to support scalable, cross-platform deployment. MCP users have 100% higher retention then non-MCP users.
- **Optimized LLM workflows in Highlight's desktop application:** implemented dynamic model routing and evaluation pipelines, developed auto-task detection via capture mechanisms, and enhanced local memory management and SLM processes to boost performance and reliability.
- **Partnered with customers and designers to drive product excellence:** facilitated user feedback sessions, translated insights into feature improvements, and aligned UX designs with technical capabilities to deliver a polished, user-centric experience users loved.

Helicone

Full Stack Engineer San Francisco, CA, USA (November 2023 - December 2023)

- Contributed to the development of Helicone, a platform for monitoring Large Language Models at scale, streamlining the management and analysis of LLMs.

Mirage

CTO Istanbul, TR (March 2023 - November 2023)

- Developed a user-friendly web application that helps clients easily create and order large numbers of synthetic images for computer vision tasks such as image classification and object detection.
- Automated the Unity Engine with an add-on for handling image generation requests, allowing real-time synthetic image testing and facilitating bulk orders exceeding 10 million images for computer vision applications.

Publications

Experimentation in Content Moderation using RWKV - Huggingface Sept 2024

- Investigated RWKV model's efficacy in content moderation, leveraging its CPU-efficient architecture
- Created a novel dataset with images, videos, sounds, and text for distillation into smaller models (558,958 for text, 83,625 for images)
- Demonstrated RWKV's potential to improve accuracy and efficiency in content moderation, paving the way for compact, resource-efficient content moderation models that understands context

Education

Flatiron School - Computer Science Manhattan, NY(Aug 2022- Nov 2022)

Isik University - Associate Degree in Computer Programming Istanbul, TR (Aug 2020– Aug 2022)