Pratik Mehta

908-215-0282 | mehtapratik@outlook.com | linkedin.com/in/pratikmehta14 | github.com/pratikmehta64

EDUCATION

University of Massachusetts Amherst — Amherst, MA
Master of Science in Computer Science — *Aug 2016 – May 2019*

D. J. Sanghvi College of Engineering, University of Mumbai — Mumbai, India
Bachelor of Engineering — *Aug 2011 – Jun 2015*

EXPERIENCE

Consultant, Machine Learning & GenAI — Nov 2020 – Jun 2024
Deloitte Digital USI — *Mumbai, India*
- Successfully designed & developed REST APIs using FastAPI, PostgreSQL, and Llama served over AWS to answer user questions about live quantitative enterprise KPI data for Beta testing by a few thousand users
- Personally initiated driving deep changes to the in-house black-box testing practice for LLM-in-the-loop applications for use across clients
- Developed use-cases and AI-adoption strategies for AI-aspirant enterprise clients

Research Engineer — Aug. 2015 – Jul. 2016
Indian Institute of Technology Bombay — *Mumbai, India*
- Trained and deployed phrase-based statistical machine translation systems for English-Marathi and English-Indonesian using medical-domain textbooks
- Designed an error analysis engine to identify and classify syntactic errors in translations, and developed a greedy algorithm for automatic post-editing, evaluated via BLEU scores and human evaluators

Research Assistant — June 2018 – Sept 2018
Information Sciences Institute, University of Southern California — *Waltham, MA*
- Studied correlation between different evaluation metrics for machine-translation systems and their downstream impact on cross-lingual search applications

PROJECTS

Sony Persona-Grounded Commonsense Reasoning challenge - API track | *AIcrowd LLM agents* — May 2025
- Developed agents and finely-tuned 'system prompts' to improve 'API and parameter choice' and function-calling accuracy
- Rated top-5 in Round 1 of this global contest that requires good performance for video-game NPC agents on both accurate function-calling and dialogue generation tasks. (Round 2 is underway)

Alfred: LLM powered dispatch of hyper-personalized notifications | *Llama* — Jan 2024
- Conceptualized and developed a proof-of-concept system to tracks users' natural-language personalization preferences based on live-data events in the application
- The system then batch-serves alerts to users based on their personal preferences for how and under which extremely precise or loosely-defined conditions they want to be notified

LegalEase: Cross-lingual semantic search engine | *Cohere, Qdrant & lablab.ai* — Mar 2023
- Developed an application that answers questions factually based on lengthy legal documents (now called Retrieval Augmented Generation)
- Additionally, the system includes a cross-lingual information retrieval system capable of answering any-language query based off of any-language document-set (compatible with about 10-12 European languages)

TECHNICAL SKILLS

Languages: Python, Rust, SQL (Postgres), HTML/CSS
Frameworks: FastAPI, gradio, streamlit, qdrant, pytorch
Developer Tools: Git, Docker, VS Code, RustRover
Libraries: pandas, NumPy, scikit-learn