sshreyv@gmail.com

(647)-675-0790

SHREY SHAH
GENERATIVE AI ENGINEER

24 Trout Run

Halifax, NS

SUMMARY

Senior AI Software Engineer with over 9 years of **software development** and 2+ years specializing in **AI agent development**. Experienced in architecting and implementing **scalable AI solutions**, including **multi-agent applications**, **RAG**, and **knowledge graph systems**. Proficient in leading AI frameworks such as **LangChain**, **LangGraph** and **MCPs.**

- designed and built **end-to-end systems** for **reliable AI interactions**, including a **message routing system** using **LangGraph** and OpenAI, and a **document management interface** with integrated **S3/OpenAI backend** and **vector store indexing** for **AI-assisted Q&A.**
- Created a framework for managing Short term and long term **memory** for the agents.
- O**ptimized and evaluated LLM agent performance** through **prompt optimization** research, **LLM-as-a-judge**, and **agent-as-a-judge frameworks**.
- **fine-tuned and quantized LLMs** and **image generation models (SDXL, Flux)** for specific tasks, achieving SOTA performance and optimizing inference using libraries like **Hugging Face transformers** and platforms like **ComfyUI**.
- Experienced in identifying and mitigating **prompt injection vulnerabilities**.
- Deep understanding of **AI-assisted coding workflows**, leveraging over 5 years of experience to significantly improve team efficiency and accelerate development cycles from **ticket to PR using agents.**
- I have holistic experience across **generative AI**, including **custom LoRa model creation**, proficiency in **ComfyUI** for **multimedia content generation**, and **voice cloning** capabilities.
- Created **MCPs** for streamlining **tool integrations** and **dynamic agent creation**.

TECHNICAL SKILLS:

- **Programming Languages:** Java, **Python**, JavaScript, Typescript
- **Knowledge of Dev-ops tools**: Docker, GitHub Actions, Harness, Kubernetes
- **Database Systems:** SQL Server, MySQL, MongoDB, Chroma
- **IDES**: IntelliJ, VScode, Cursor
- **GenAI:** Langchain, Langgraph, Langsmith, W&B, HuggingFace, ComfyUI

WORK EXPERIENCE

Senior AI Software Engineer   **Vivun**   **May 2023 – Present**
Environment: Langgraph, Python, fastapi, nextjs, LLMS, RAG, Cursor

- Became an AI advocate within the team and organization. Led **training sessions** for **Developers** on integrating and using **AI**.
- Always researched and created **POCs** about the latest advancements in **LLMs**, **fine-tuning techniques**, and **AI tools** and shared findings with the teams as new advancements come up. Rapidly **proto-typed research projects**, which allowed us to explore different options and we were able to make better decisions because of that.
- Led the development of **multi-agent applications** using **LangGraph** for efficient **LLM** and **RAG-based systems**.
- Developed **POCs** exploring **RAG solutions**, including **Hybrid RAG**, **GraphRag** and **Agentic RAG**. Integrated **long-term and short-term memory** in the agents using **knowledge graphs** and used efficient **rag** for retrieval.
- Developed **CI/CD pipelines** for evaluating, and deployment of the agents.
- Developed an **instruction-tuned dataset** and used it to **fine-tune models** for the purpose of using them for **LLM as a judge** and **Agent as a judge techniques** for **evaluating performance**.
- Executed comprehensive **adversarial testing** on the chatbot, employing various **prompt injection methodologies**.
- Used **MCP tools** to create a **ticket to PR workflows** in **Cursor**, to release features really really fast. Developed **MCPs** for **dynamic agents creation** and dramatically improved the **speed to production**.
- Led **AI assisted coding** efforts org wide, Coached the entire dev team on **efficient workflows**. Automated **prompt engineering** for agents by creating **evaluation frameworks** for them first and then ran **experiments at scale** to find out the best performing solution, which allowed us to move really fast without compromising on quality.
- Leveraged **Langsmith** for **monitoring**, **evaluating** and **optimizing** the agents and prompts.
- Heavily influenced the **UI/UX** of the agents based on my findings by doing **rapid proto-typing**. Stay up to date on trends in **LLM architectures**, **agent frameworks**, **evaluation strategies**, and **GenAI standards**.

Cursor Ambassador   **Anysphere**   **Mar 2025 - Present**

- Helping bridge the gap between the AI community and Cursor team
- Hosted webinars & talks on Maven resulting in thousands of individuals learning more about Cursor Community
- Engaged and supported Cursor AI users resulting in meaningful changes prioritized by the Cursor team.
- 2000+ hrs of cursor AI assisted coding experience resulting in building full stack apps within an hour.

Lead SDET   **Priceline**   **June 2021 – May 2023**
Environment: Agile, Mocha, Chai, Postman, WebDriverIO, Axios, GRPC, Jira, Node.js, SQL, JavaScript, Splunk, GraphQL
- Led development of an end-to-end test automation framework using wdio, Appium, Mocha and Saucelabs for web, mobile and API testing, significantly improving test coverage and reducing regression testing time. This allowed new features to be released faster, resulting in increased customer conversion rate.
- Implemented GitHub actions for automating the entire release process, including PR creation and merging, as well as confluence release documentation, reducing release time.
- Developed backend for one of core products using Spring Boot, supporting high volume of daily transactions and driving strong revenue growth after launch.
- Created algorithms to auto-generate tests from historical test-data, reducing manual test creation and maintenance efforts. This allowed the testing team to keep pace with aggressive development timelines, while improving test coverage and quality.
- Integrated test automation framework with Zephyr Scale to provide end-to-end traceability. This improved visibility into overall test coverage and helped identify areas needing more test focus, reducing defects.
- Implemented Splunk dashboards and alerts for real-time application performance monitoring. This enabled pre-emptive identification and resolution of issues, improving uptime and customer satisfaction.
- Provided technical leadership and mentoring on test automation best practices using Mocha, WebDriverIO and JavaScript, improving team productivity and efficiency.
- Played a pivotal role in the recruitment process for SDETs conducting interviews to match job requirements.

Automation Developer   **Powerhub Inc.**   **sept 2020 – June 2021**
Environment: Agile, Mocha, Chai, Postman, Puppeteer, TFS, Azure Dev-ops, Node.js, JIRA, MongoDB, JavaScript
- Developed an innovative test automation framework using Cypress, Mocha, Chai, and JavaScript from scratch, improving test automation efficiency.
- Implemented Eslint and SonarQube tools to consistently maintain code quality, decreasing bugs and errors.
- Employed Postman, Mocha, Chai, JavaScript for API testing, and JMeter for UI/API performance testing,
- Crafted MongoDB queries in Roto3T for database validation against the UI, enhancing data accuracy..

QA automation developer   **Parsedata**   **Apr 2020-sept 2020**
Environment: Agile, Azure, Linux, Microsoft Teams, Azure AD, Bash scripting, Selenium, python, Java, GitHub, SoapUI
- Gained hands-on experience with Microsoft Azure DevOps portal and Active Directory.
- Developed a robust testing framework using Python, Selenium, Pytest, and POM, reducing bug reports and saving time.

Quality Assurance Automation Engineer   **NXM Labs INC**   **Feb 2019-Apr 2020**
Environment: Agile, TDD, Selenium WebDriver, python, JAVA TestNG, JIRA, IoT, POM Design Pattern, GitLab, SoapUI, JMeter, ADB, **Appium, Gherkin**
- Led comprehensive testing of applications, including Blockchain, Mobile & Web apps, SQL databases, and IoT devices, ensuring product quality and enhancing user experience.
- Developed a data-driven framework for API, Web, and Mobile app testing, improving testing accuracy and speed.
- Engineered an API testing framework using Python and Pytest for IoT devices, increasing testing efficiency.
- Conducted load testing on APIs using Apache JMeter and Artillery, optimizing application performance and reliability.
- Performed code quality analysis using Sonarlint and SonarQube, reducing errors in the QA framework.

Software Tester   **Webyoutechnologies**   **May 2014 -Apr 2016**
- Developed Selenium test cases using the **Hybrid framework** (Data Driven Framework with POM and TestNG)
- Validated the scripts to make sure they have been executed correctly and meet the scenario description. .

EDUCATION

Master of Engineering   **University of Windsor**   **May 2016- Aug 2017**

Bachelor of Engineering   **Gujarat Technological University**   **Aug 2011-May 2015**