Shriyansh Singh

+1-930-333-5141 | **shriyansh.singh24@gmail.com** | **LinkedIn**

SUMMARY

Machine Learning Engineer with expertise in **LLM application development** and **red teaming**. Skilled in developing robust **prompt engineering** techniques and identifying **AI vulnerabilities**

PROFESSIONAL EXPERIENCE

AI Security Research Intern — *April 2024 - Dec 2024*

Hyphenova AI — *Los Angeles, CA*

- **Developed** a systematic **LLM vulnerability testing framework** that identified 14 novel jailbreak patterns across GPT and Claude models, resulting in direct model safety improvements
- **Engineered Python scripts** leveraging OpenAI, Anthropic, and Cohere APIs to implement adaptive prompt testing that automated discovery of model guardrail weaknesses
- **Created** an automated prompt evaluation system measuring ROUGE, BLEU, and custom risk metrics that reduced manual review time by 85% while increasing detection precision
- **Designed** and deployed a **red team training program** for 25+ engineers, increasing vulnerability discovery rates by 62% through structured adversarial techniques
- **Presented** vulnerability findings at two internal security conferences, leading to the adoption of 8 new guardrail implementations for production models

Machine Learning Research Assistant — *May 2022 - Oct 2022*

Enterprise Business Technologies Pvt Ltd — *Mumbai, India*

- **Built** a **LLM-based data generation system** in **TypeScript** and Python that produced synthetic training datasets, improving downstream model performance by 27% on rare edge cases
- **Implemented** prompt engineering techniques for extracting structured information from unstructured text, achieving 91% accuracy on complex document parsing tasks
- **Conducted** systematic evaluation of frontier LLMs (GPT-3.5/4, Claude) across 8 performance dimensions, creating a comprehensive visualization dashboard for model selection
- **Developed** a custom data quality assessment framework using **frontier LLMs as evaluation tools**, reducing annotation costs by $45K while maintaining 97% quality standards

PROJECTS

LLM Vulnerability Scanner | *Python, LangChain, API Integration, Statistical Analysis* — *Oct 2024 – Jan 2025*

- **Engineered** a comprehensive **red teaming tool** that systematically probes model boundaries through 1000+ parameterized attack vectors, identifying critical safety vulnerabilities
- **Implemented** statistical analysis of model responses using custom metrics and established benchmarks (MAUVE, ROUGE) to quantify vulnerability severity
- **Developed** an automated reporting system that documented successful attack patterns and suggested mitigation strategies for model providers

Synthetic Data Generator for LLM Fine-tuning | *Python, TypeScript, AWS, LLM APIs* — *May 2024 – Dec 2024*

- **Created** a **synthetic data pipeline** using frontier LLMs that generated customized training examples for specialized domains, improving downstream model performance by 35%
- **Built** a web interface with **TypeScript** that allowed non-technical users to define data generation parameters and quality criteria through intuitive prompting patterns
- **Deployed** the system on **AWS Lambda** with S3 integration, enabling scalable generation of training datasets with automated quality filtering and format validation
- **Implemented** a sophisticated data evaluation framework using statistical measures and LLM-based assessors to ensure synthetic data maintained high diversity and realism

SKILLS

Programming: Python, **TypeScript**, LangChain, HuggingFace Transformers, SQL
ML/LLMs: Prompt Engineering, **Red Teaming**, **Model Evaluation**, GPT API, Claude API, Jailbreak Detection, RLHF
Evaluation: ROUGE, **BLEU**, **MAUVE**, Perplexity, Statistical Analysis, A/B Testing
Tools & Platforms: AWS (Lambda, S3, SageMaker), Git, Docker, Jupyter, VS Code

EDUCATION

Indiana University Bloomington — Aug 2023 – May 2025

Master of Science in Data Science — *Indiana, United States*