Contact Information

Name: John Doe
Email: john.doe@example.com
Phone: +1 (555) 123-4567
LinkedIn: linkedin.com/in/johndoe
GitHub: github.com/johndoe
Location: Mountain View, CA

Education

Ph.D. in Computer Science, Specialization in Machine Learning — 2018
Stanford University — Stanford, CA
GPA: 3.9/4.0

M.S. in Data Science — 2015
Carnegie Mellon University — Pittsburgh, PA

B.S. in Mathematics and Statistics — 2013
University of California, Berkeley — Berkeley, CA

Professional Experience

Senior AI Safety Data Scientist — 2021 – Present
Tech Innovations Inc. — San Francisco, CA

- Led cross-functional initiatives for **AI safety** and **risk mitigation** in **Generative AI** products, reducing critical safety incidents by **18%** through advanced **machine learning techniques** and **statistical methods**.
- Developed and deployed **scalable safety solutions** for **Large Language Models (LLMs)**, focusing on **harmful content moderation**, **child safety**, and **imminent threat analysis** across diverse datasets exceeding **10TB**.
- Managed end-to-end projects, defining **project scope**, **goals**, and **deliverables** for **automated data pipelines** and **self-service dashboards**, providing timely **insights** to executive leadership and improving data accessibility by **30%**.
- Mentored **5 junior team members** on best practices in **large-scale data analysis**, **data science methods**, and ethical considerations in **AI/LLM-powered solutions**, fostering a culture of **proactive communication** and **collaboration**.

Data Scientist, Trust & Safety — 2018 – 2021
Global Social Platform — Seattle, WA

- Applied **advanced machine learning techniques** to identify trends and patterns in **abuse and fraud disciplines**, enhancing detection accuracy by **25%** and reducing false positives by **15%** across a user base of **500M+**.
- Conducted in-depth **threat analysis** and **web security** assessments, developing actionable **insights** from **quantitative** and **qualitative data** to strengthen platform protection measures and inform strategic decision-making.
- Utilized **SQL** for large-scale data extraction and analysis, and **Python** for developing **scripting languages** and custom algorithms to automate data processing workflows, saving over **200 hours** annually.
- Collaborated with engineering teams to examine existing protection measures, uncovering potential shortcomings, and recommending **security enhancement** strategies that improved system resilience.

Projects

LLM Safety & Bias Mitigation Framework 2023
GitHub Link

- Developed a comprehensive framework for evaluating and mitigating biases in **Large Language Models (LLMs)** using **Python**, **Julia**, and advanced **statistical methods**.
- Implemented novel **prompt engineering** techniques and **fine-tuning LLMs** strategies on custom datasets to reduce harmful outputs and improve model fairness by **12%**.
- Conducted extensive **data analysis** to identify subtle trends in model behavior, generating summary statistics and drawing critical **insights** on potential real-world safety risks.
- Explored applications of **Vertex AI** and foundational models like **Gemini** for enhanced safety evaluation and developed automated reporting for continuous monitoring.

Automated Harmful Content Detection System 2022
GitHub Link

- Designed and deployed an **AI/ML-powered solution** for real-time detection of harmful content, processing over **1 million** data points daily with **95%** accuracy.
- Leveraged **C++** for performance-critical components and **Python** for model training and data orchestration, ensuring **scalable solutions** for high-throughput data streams.
- Developed **automated data pipelines** and integrated with existing security infrastructure to provide immediate alerts for **imminent threat analysis**, significantly reducing response time.
- Focused on **security enhancement** by continuously examining protection measures and integrating feedback loops from human moderators to refine model performance and adapt to evolving threats.

Technical Skills

Programming Languages: Python, C, C++, Julia, SQL, Scripting languages
Frameworks & Libraries: Machine Learning (ML), Artificial Intelligence (AI), Generative AI, Large Language Models (LLMs), AI/LLM-powered solutions, Vertex AI, Foundational models (Gemini, Juno, Veo), TensorFlow, PyTorch, Scikit-learn, Pandas, NumPy
Tools & Technologies: Data analysis, Identifying trends, Generating summary statistics, Drawing insights (quantitative data, qualitative data), Project management (defining project scope, goals, deliverables), Large-scale data analysis, Data science, Abuse and fraud disciplines, Web security, Harmful content moderation, Threat analysis, Prompt engineering, Fine-tuning LLMs, Risk mitigation (generative AI), Real-world safety (LLM/AI technology), Imminent threat analysis, Child safety, Developing scalable safety solutions, AI products, Advanced machine learning techniques, Advanced AI techniques, Statistical methods, Data science methods, Examining protection measures, Uncovering potential shortcomings, Developing actionable insights, Security enhancement, Developing automated data pipelines, Self-service dashboards, Providing timely insights, Driving progress, Solving problems, Mentoring (junior team members), Technical know-how, Google Cloud Platform (GCP)
Databases: SQL, NoSQL (e.g., MongoDB, Cassandra)
Cloud Platforms: Google Cloud Platform (GCP), AWS, Azure