Harmandeep Kaur

✉ harmandeep.kaur5189@gmail.com | 669-237-0009 | Dublin, CA

https://www.linkedin.com/in/kaurharmandeepbains/

Experience

PRYON — ***San Francisco, CA***
Senior Machine Learning Engineer — ***Feb 2023 to Present***

- Created chatbots using Large Language Models (LLMs), integrating multi-modal features with Open AI and LLAMA, and enhancing user interfaces with Streamlit for better interaction.
- Architected advanced chatbots leveraging Langchain, RAG, and Vector DB embeddings, improving conversational AI with enhanced language understanding and context awareness.
- Utilized Ray for scalable chatbot processing and deployment, seamlessly integrating multi-modal functionalities and enhancing user interaction with sophisticated language models.
- Applied Huggingface models to boost conversational AI capabilities with cutting-edge natural language understanding and generation.
- Developed machine learning workflows in Databricks notebooks, optimizing data processing and model deployment pipelines.
- Streamlined machine learning model serving APIs with FastAPI, Spark, and PySpark on Databricks, ensuring real-time inference and integration with PostgreSQL, ElasticSearch, and S3 databases.
- Enhanced chatbots by integrating LLAMA2 for text-to-SQL capabilities, facilitating the conversion of unstructured text data into structured SQL queries for advanced data analysis.
- Developed data analysis and visualization chains using LLAMA2, enabling efficient analysis and plotting of CSV data through automated workflows and agent-based systems.
- Managed code repositories using Git and GitLab, ensuring collaborative development and adherence to version control best practices.
- Maintained efficient CI/CD pipelines with EKS, Helm, and ArgoCD, facilitating seamless ML model deployment to production environments.
- Collaborated with data scientists and engineers to deploy models in production, establish monitoring strategies, and track key performance indicators (KPIs) for continuous improvement.

Cisco Systems, Inc. — ***San Jose, CA***
Senior Machine Learning Engineer — ***Jan 2021 to Jan 2023***

- Developed and trained sophisticated machine learning models using PyTorch, emphasizing convolutional and recurrent neural networks. Leveraged PyTorch Lightning to optimize training workflows and utilized Weights & Biases for comprehensive experiment tracking and visualization, ensuring clear and reproducible outcomes.
- Deployed machine learning models with PyTorch, focusing on deep learning structures. Utilized PyTorch Lightning to modularize and scale code, while FastAI accelerated prototyping for efficient experimentation and development.
- Automated hyperparameter optimization using Scikit-Optimize, markedly enhancing model accuracy and efficiency across various scenarios.
- Leveraged MLflow for experiment management and result tracking, ensuring that models met specified performance benchmarks prior to deployment.
- Fine-tuned deep learning models with transfer learning and Hugging Face Transformers, adjusting architectures and parameters based on validation outcomes. Further refined models using FastAI.
- Deployed models into production environments with Docker for containerization and FastAPI for robust API development, facilitating scalable and efficient model serving. Employed TorchServe for the deployment and management of PyTorch models, ensuring high-performance inference.

- Established monitoring and management frameworks for models using Prometheus and Grafana for real-time performance tracking. Integrated Evidently AI to monitor data drift and model performance degradation, enabling proactive maintenance and updates.
- Collaborated with cross-functional teams utilizing Git for version control and GitHub for code sharing and review. Automated testing and deployment processes with CI/CD pipelines using GitHub Actions and GitLab CI, enhancing development efficiency and minimizing manual errors.
- Managed cloud-based machine learning workflows on AWS, orchestrating with Kubernetes and Kubeflow. Automated deployment and scaling using Terraform and Apache Airflow, ensuring consistent and reliable model training and serving across environments.
- Applied ensemble learning techniques, including bagging, boosting, and stacking, to enhance model robustness and performance by integrating diverse models.
- Employed data manipulation and analysis tools such as pandas and NumPy for preprocessing and cleaning datasets, ensuring accurate and reliable model training.
- Conducted exploratory data analysis (EDA) using Matplotlib and Seaborn for visualization, identifying patterns and insights to guide feature engineering and model development.

Labcorp ***Burlington, NC***
Data Engineer ***May 2019 to Dec 2020***

- Developed deep learning models using TensorFlow and Keras, employing TensorFlow Extended (TFX) for scalable end-to-end ML pipelines and optimizing model performance through TensorBoard analysis.
- Integrated TensorFlow Serving and TensorFlow Lite for efficient deployment of ML models on various platforms, ensuring smooth inference across diverse hardware environments.
- Leveraged TensorFlow Model Garden to access pre-trained models, accelerating development cycles and enhancing model robustness and efficiency.
- Employed AutoML techniques within TensorFlow to automate model selection, hyperparameter tuning, and architecture search, streamlining the model development process for improved productivity.
- Curated and managed a repository of reusable machine learning components on TensorFlow Hub, facilitating prototyping and collaboration among cross-functional teams on cloud platforms.
- Utilized Docker-based containerization for ML applications, ensuring consistent and reproducible deployment environments across development, testing, and production stages.
- Deployed ML models into production environments using continuous integration and delivery pipelines, enabling seamless updates and monitoring with Kubernetes.
- Implemented scalable and efficient data preprocessing pipelines, incorporating feature engineering techniques and data augmentation to enhance model performance and robustness.
- Collaborated closely with data scientists, software engineers, and domain experts to translate business requirements into ML solutions.
- Actively participated in code reviews, knowledge-sharing sessions, and cross-functional collaborations, fostering a culture of continuous learning and improvement within the ML engineering team.

THOMSON REUTERS ***San Francisco, CA***
Data Analyst ***Aug 2017 to April 2019***

- Executed data collection through SQL queries and Python scripts, utilizing web scraping tools like BeautifulSoup and Scrapy to gather data from various sources, including APIs and direct database access.
- Engineered data ingestion pipelines using Apache Kafka and Apache NiFi to handle real-time data streams and batch data loads, ensuring smooth data flow into analysis frameworks.
- Developed SQL scripts for data extraction and conducted data cleaning and transformation with Pandas and Dask in Python, processing both structured and unstructured data from PostgreSQL and MySQL databases.
- Employed Excel functions and Python libraries such as Pandas and NumPy for exploratory data analysis (EDA), identifying patterns, anomalies, and correlations in large datasets to inform further analysis.
- Created dynamic data visualizations using Matplotlib, Seaborn, and Plotly in Python, and developed interactive dashboards with Tableau and Power BI for effective data exploration and stakeholder presentations.
- Performed statistical analysis using Python libraries such as SciPy and StatsModels, conducting hypothesis tests, regression analysis, ANOVA, and time-series forecasting to support data-driven decision-making.
- Enhanced data storage and retrieval by designing optimized database schemas in PostgreSQL and MySQL.
- Automated repetitive data processing tasks by building ETL pipelines with Apache Airflow and Talend, improving workflow efficiency and reliability in data transformation processes.

- Compiled and presented detailed analytical reports using advanced features of Microsoft Excel and Google Sheets to communicate key findings and recommendations to stakeholders effectively.

Wipro ***India***

Software Engineer ***Aug 2011 to Jan 2016***

- Provided efficient support in C#, ASP.NET, Java, and SQL, ensuring quick resolution of user issues.
- Collaborated closely with engineering teams using ASP.NET, and Java to analyze and address user queries, identify technical challenges, and deliver effective solutions.
- Maintained detailed documentation of user interactions, technical issues, and resolutions provided using Documentum.
- Participated in process improvement initiatives, analyzing support tickets in SQL to enhance the overall user experience.
- Successfully resolved user queries and technical issues through effective communication and strong problem-solving skills in C#, ASP.NET, Java, and SQL.
- Worked with engineering teams using C#, ASP.NET, and Java to provide timely and accurate solutions to user problems.
- Maintained standard operating procedures, analyzed support tickets for process improvements using SQL, and ensured highquality user experiences.
- Conducted training sessions and supported user feedback projects, demonstrating excellent problem-solving abilities in C#, ASP.NET, Java, Documentum, and SQL.

Education:

International Technological University **San Jose**
Masters of Computer Science **2016-2019**

Technical Skills

Programming Languages:	Python, SQL
ML Frameworks/Libraries:	Pytorch, Huggingface, Keras/Tensorflow, MLFlow
Cloud Platforms:	AWS Bedrock, AWS Sagemaker, EKS (Ray Clustering)
Data Processing and Analysis:	Numpy, Pandas, Spark, Pyspark, Databricks
Natural Language Processing (NLP):	Huggingface, Pytorch, NLP, Open AI/LLAMA/Anthropic
Web Development and Deployment:	Streamlit, FASTAPI, Docker, Kubernetes, Helm
Other Tools and Technologies:	LLM, Gen AI, RAG, Vector DB, Langchain, Pydantic, Pytest