Pavan Chhatpar

Atlanta, GA | pavanchhatpar@gmail.com | (857) 930-1785 | linkedin.com/in/pavan-chhatpar | pavanchhatpar.github.io

TECHNICAL KNOWLEDGE

Languages:	Python, C++, Java, Julia, Node.js, TypeScript, SQL
ML & AI Frameworks:	PyTorch, TensorFlow, Transformers, TRL, PEFT, XGBoost, sklearn, LangChain, LangGraph, NumPy
MLOps & Pipelines:	MLflow, TensorBoard, Spark, Databricks, KServe, Kubernetes, Docker, Airflow, GitHub Actions
Databases:	Postgres, Milvus, Dremio, Hive, Vertica, MongoDB, MySQL, MS SQL, Oracle, SQLite

PROFESSIONAL EXPERIENCE

Honeywell International Inc, Atlanta, GA

Sr. Advanced AI Engineer Feb 2025 – Present

- Deployed **open-source LLMs** on **Kubernetes** using **KServe** and **vLLM**, delivering **50% cost savings** while maintaining **<5% loss in accuracy** for agentic applications and ensuring **data compliance** in restricted cloud regions
- **Architected and mentored** the development of **guardrails and red-teaming agents**, establishing rigorous validation protocols to **prevent agent misuse** in production
- Built a scalable **Agent evaluation framework** utilizing **LLM-as-a-judge, retrieval, and ranking metrics** to automate benchmarks, accelerating deployment for **~10 business-specific chatbots**
- Spearheaded **Edge GenAI POCs** on NVIDIA Jetson for **Honeywell's Board of Directors**, utilizing **quantized LLMs** and **local vector databases** to enable private on-device intelligence
- **Led design and implementation** of **Model Context Protocol** (MCP) servers enabling **IoT device control** for autonomous, efficient operations for buildings with agentic apps; built in Python and containerized to deploy on NVIDIA Jetson devices
- **Led fine-tuning of quantized LLMs** using **LoRA** and **GRPO**, producing domain-specialized models with lesser prompt token overhead **reducing latency by 45%** at the expense of **10% more memory from LoRA weights**
- **Architected vectorization pipelines in PySpark** that automate tasks from ingestion through quality assessment, delivering **~25M vectors** of up-to-date data for consumption in agentic applications
- **Mentored the team** in their journey to transition from academia to industry-grade ML engineering, **fostering best practices** in every step of ML systems' lifecycle

Advanced Data Scientist Jan 2022 – Feb 2025

- **Co-architected** Honeywell's Agentic AI platform, built on LangChain, enabling **low-code agentic orchestration and workflows** that empower business units to focus on data preparation and quality assessment
- Led the deployment of autoscaling APIs on **Kubernetes** for agentic workflows that leverage **ReAct and Chain of Thought reasoning with LLMs**; the system supports **1000+ questions per minute** maintaining accuracy between 80-85%
- Designed platform capabilities for prompt engineering, API-as-a-tool usage, and dynamic LLM selection
- Trained **multi-modal transformers** for classifying IoT sensor data composed of **text and time series features**, using distributed training over multiple GPUs in a Databricks cluster with **PyTorch, DeepSpeed, and MLflow**
- Built a scalable **few-shot learning** feedback loop on a **real-time inference API** for the multi-modal transformer model, using Redis, FastAPI, and Kubernetes; it reduced digital twin onboarding time from **3 weeks to ~7 hours**
- Developed a **patent search engine** that uses PySpark to scale as a **distributed index** over a partitioned parquet dataset using **FAISS**, which achieved query speeds between 50-70ms with the help of PySpark's caching capabilities

Data Scientist II Jan 2021 – Jan 2022

- Implemented **anomaly detection frameworks** to identify procurement fraud by combining **unsupervised learning** and **rule-based analytics**, collaborating closely with finance and supply chain teams
- Developed a **feedback-driven prioritization algorithm** that ranked alerts based on reviewer interactions, **reducing manual review workload by ~65%**
- Boosted Forge Insights' platform adoption and efficiency by **streamlining data access** with fsspec and enforcing robust security through Privacera, resulting in **~90% reduction** in I/O speed

Data Science Intern Jun 2020 – Aug 2020

- Applied **Named Entity Recognition (NER)** to extract key business entities from supplier contracts
- Fine-tuned **transformer-based NLP models** for contract clause classification, enabling compliance monitoring
- Scaled the **contract processing pipeline** using **PySpark** to handle large document batches on a recurring schedule
- Researched methods to **link the contract clauses with transactional data**, providing foundational work for risk models

Wayfair, Boston, MA May 2019 – Dec 2019

Data Science Co-op

- Trained **Survival Analysis Models** on large-scale time-series data using **recurrent neural networks in Python**
- Developed data pipelines using **PySpark** for data from **Hive**; scheduled daily jobs to run them in **AirFlow**
- Conducted **backtesting** to evaluate model efficacy over a variety of time horizons
- Engaged in **stakeholder meetings** to leverage their domain knowledge in **feature engineering**

dotin, Fremont, CA (*Remote*) Mar 2018 – Jun 2018

Software Engineer Intern – Machine Learning

- Developed ML training, testing, and predictor modules with pipelining using **Python**, **Julia**, and **Java**
- Contributed to **feature engineering** tasks in the ML training pipeline
- Contributed to maintaining data collection through **Amazon Mechanical Turk**

PROJECTS & PUBLICATIONS

Deep Question Generation on SQuAD dataset Apr 2020
Master's Project, Northeastern University, Boston, MA

- Developed a **generative deep neural network**, using Tensorflow, for question generation from paragraphs
- Optimized convergence using Copy Mechanism, such that the generated questions could get answers with an F1 score only **18% lower** than the original questions
- Simplified the model architecture with Copy mechanism, enabling a **lightweight GRU-Attention model** for efficient sequence learning with a thin vocabulary size
- **Leveraged Beam Search** for better exploration of top k output tokens at each step, increasing the robustness of the generated question compared to a greedy approach of using just the top token at each decoding step
- Contributed a generic CopyNet TensorFlow implementation as an **open-source package** via GitHub

The precision of case difficulty and referral decisions: an innovative automated approach Aug 2019
Nair Hospital and Dental College, Mumbai, India

- Collaborated on an **ML solution** with a team of dentists to **predict the difficulty of an Endodontic case** before treatment using TensorFlow and sklearn, achieving a sensitivity score of **94.96%**
- Published in Clinical Oral Investigations, Springer (**Impact Factor - 3.3**)
- Developed an Android app leveraging TFLite (now LiteRT) for on-device inference, enabling real-world application and validation of the research findings

Vehicular Traffic Abatement May 2018
Final year Project, University of Mumbai, Mumbai, India

- Developed a solution to vehicular traffic using **neural networks** in a team of four, facilitating users with the prediction of vehicular traffic based on time and location, with an accuracy of **90.73%**
- Published the project work as two phases in **IEEE**, Nov. 2018 and in **IJRASET** Volume 6, Jul 2018

EDUCATION

Master of Science in Computer Science – Northeastern University, Boston, MA Sep 2018 – Dec 2020
Khoury College of Computer Sciences | **GPA: 4.0/4.0**

Bachelor of Engineering in Computer Engineering – University of Mumbai, India Jul 2014 – May 2018
Vivekanand Education Society's Institute of Technology | **GPA: 8.99/10.0**