Pengwei Wang

Austin, TX
(301) 525-9881
pwwang08@outlook.com

SUMMARY

Innovative AI/ML professional with 3+ years of experience building and scaling ML and Generative AI applications. Specialized in breaking down business problems and translating technical solutions into strategic outcomes. Proven record in collaborating across cross-functional stakeholders to solve dynamic challenges in both Big 4 and startup companies. Interested in the integration of advanced retrieval systems and autonomous agents to transform enterprise productivity.

TECHNICAL SKILLS

AI/ML Techniques: NLP, Neural Networks, Transformers, Semantic Search, Vector Embeddings, MLOps, LLMOps, Agentic AI

ML Frameworks: PyTorch, TensorFlow, Scikit-Learn, XGBoost, Pandas, LangChain, LlamaIndex, Hugging Face, MLflow

Cloud & Dev: Azure AI Services, Google GCP Vertex, Databricks, Spark, MilvusDB, MongoDB, FastAPI, Docker, Git, Pydantic

Languages: Python (Proficient), SQL (Proficient), R (Familiar)

PROFESSIONAL EXPERIENCE

Deloitte *July 2023 – December 2024*
AI Automation Analyst

- Enhanced LLMOps workflows leveraging Databricks to process model logs and compute evaluation metrics. Integrated LLM-as-a-Judge and multi-class scoring to maintain relevance and factuality above 85% in production
- Identified security vulnerabilities in the internal AI assistant's research and citation processes. Implemented query classification using BERT with NLP content filtering guardrails in Python that reduced data leakages by 10%
- Designed POCs for strategic clients in collaboration with engineers and product owners, translating business requirements into technical specifications. Delivered pilot proposals for Visa chatbot and Costco returns analytics product.
- Hosted knowledge sessions to address GenAI risks, vector retrieval challenges and data extraction frameworks. Created a standardized data extraction process adopted across six client engagement teams
- ***Project Highlight: GenAI Legal Document Q&A System***
    - Led the development of a GTM product to automate document review at scale, streamlining assessment reports, interactive document chat and quality review to save $62K+ of manual efforts in FY25
    - Architected custom Retrieval Augmented Generation (RAG) system using OpenAI with Document Intelligence, deployed in production workflows. Optimized retrieval efficiency by 7x using vector database
    - Enhanced end-to-end accuracy by 12% with hybrid search, reranking, parameter tuning and few-shot prompting
    - Collaborated with 30+ SMEs and end users to systematically fine-tune model output, integrating compliance specific terminologies with user feedback to minimize edge case failures

Burlington Stores *Jan 2023 – May 2023*
Data Scientist Intern

- Engineered a scalable ETL pipeline with PySpark to enable 5x faster processing on 70M+ records of raw POS data.
- Deployed multiple regression and XGBoost models in Python to quantify the impact of shrinkage and damaged items
- Created Tableau dashboards to highlight the top shrinkage drivers and visualize discrepancies across 1100+ stores
- Presented strategic insights to 200+ non-technical retail professionals at a retail conference (RILA), leading to 15+ executive follow-ups on best practices and strategies to mitigate shrinkage

PKWG Ltd. *Jan 2022 – June 2022*
Supply Chain Data Analyst

PROJECT

DeepAlpha – Unified AI Investment Intelligence | Claude, MCP, Pydantic, FastAPI, Flask

- Developed a multi-agent system to automate real-time earnings analysis, utilizing Model Context Protocol (MCP) to provide specialized agents with tools like web browsing and data parsing that accelerated the decision-making process by 70%
- Built an NLP pipeline analyzing news, YouTube videos and insider trading data to generate personalized portfolio reports
- Integrated macro data from financial APIs with technical chart patterns into a multimodal LLM, delivering buy, sell and hold recommendations via a Plotly interface that achieved 85% success rate over two months

EDUCATION

M.S. in Business Analytics (STEM), The University of Texas at Austin *May 2023*

B.A. in Economics, Syracuse University *December 2020*

CERTIFICATION

Advanced AI Practitioner, DataCamp *December 2024*

Generative AI Data Scientist, Advanced Prompt Engineer, Deloitte AI Academy *October 2024*

Certified ScrumMaster, Scrum Alliance *December 2023*