Patrick McBride

linkedin.com/in/pmmcbride | github.com/pmcbride | 530.320.8475 | pmmcbride@gmail.com | Bay Area & Sacramento, CA

TECHNICAL SKILLS

- **Core Programming:** Python, C++
- **Machine Learning Frameworks:** PyTorch, TensorFlow, Keras, Scikit-learn, NumPy, Pandas
- **Cloud and DevOps:** Docker, FastAPI, AWS S3, AWS ECS, GCP, Kubernetes, Linux
- **Data Management:** SQL, MongoDB, Redis, Neo4j, Milvus
- **Large Language Models:** Chatbots (OpenAI API, Agents, RAG, GraphRAG, Function Calling, Structured Outputs, Prompt Engineering), Huggingface (Transformers, Accelerate, Datasets), Fine-tuning techniques (PEFT, QLoRA, DPO), Speech-to-text (Whisper, Word2Vec, PyAnnote, Speaker Diarization)
- **Deep Learning:** GAN (SRGAN, Pix2Pix), Autoencoder, LSTM, GRU, VAE, ResNet, RNN, Reinforcement Learning
- **Machine Learning:** XGBoost, LightGBM, Time Series Analysis, Recommendation Systems

WORK EXPERIENCE

Verizon | Staff AI Software Engineer | Remote, CA Sept 2022 - Present

- Spearheaded the development of a comprehensive Verizon contracts AI Chatbot and preprocessing pipeline, achieving 90% accuracy in detecting and summarizing deviations from the Verizon standard terms found in contract Master Agreements while accounting for all Amendments.
- Directed a team of one other engineer in creating a Neo4j graph-based AI tool for long contracts, which automates the process of parsing Verizon Contract Master Agreements and Amendments into sections, identifies the relationships between specific Amendment sections and Master Agreement sections, and populates the Neo4j graph for GraphRAG, thereby reducing hallucinations.
- Fine-tuned open-source Llama and Mistral models for insurance language extraction from long contracts, achieving 90% accuracy and saving 5000+ hours/quarter of third-party labor.
- Fine-tuned open-source BERT-based models for long document classification, automating 400+ hours/month of engineering labor.
- Created an on-prem GenAI competitive intel chatbot capable of answering data-driven queries by making SQL queries to Google BigQuery and analyzing the data before providing summarized responses and trends. Developed POC in 2 weeks with MVP and presentation to executives ready in 1 month.
- Developed real-time speech-to-text transcription for agent-customer calling transcripts using OpenAI Whisper, reducing the delay time from 30 seconds to 5 seconds.

ApplyPass | Head of AI & Machine Learning | Remote, CA | applypass.com June 2023 - Present

- Built and deployed the entire ApplyPass AI backend with OpenAI API in Python and FastAPI, handling 800K requests/month and 3B tokens/month.
- Created the ApplyPass Job Classifier using OpenAI API's function calling to convert job listings into JSON format, matching the OpenAPI spec schema in our SQL backend database.
- Developed the ApplyPass Answer Generator that produces tailored responses for job application questions based on user resumes and other personal data.
- Migrated initial MVP model from GPT-4 to GPT-3.5, achieving comparable accuracy at 5% of the original cost.
- Fine-tuned a custom GPT-3.5 model to solve a customer job filter issue, improving job filter accuracy from 80% to 95%.
- Developed and implemented a content-based recommendation system using OpenAI's embeddings API to evaluate the semantic similarity between user resumes and job listings, resulting in an improvement in cold-start recommendation accuracy from 30% to 70%.
- Designed and implemented a recommendation system using a Two-Tower Model trained on user feedback to boost recommendation accuracy from 70% to over 90%.

KLA | Senior AI Software Engineer (Data Science for Deep Learning & AI) | Milpitas, CA Mar 2020 - Sep 2022

- Developed a data storage API integrating PostgreSQL ORM with MinIO object storage, improving write speed by 4X and read speed by 6X for image datasets.
- Created a multi-container application using Docker Compose for full-image DL inference in Python, analyzing 10,000+ images per workload with TensorFlow, Scikit-learn, and OpenCV.
- Acted as the primary Data and Systems Engineer for N+2 product shipment, developing data collection and product workflow, and training end users before alpha shipment.
- Developed a Python API interface to a cross-division RESTful API for large-scale image and metadata analysis.

KLA | Senior AI Data Scientist | Milpitas, CA Mar 2016 - Mar 2020

- Characterized deep learning image classification techniques based on GoogLeNet, transforming workloads from 10 hours to less than 30 minutes and optimizing training accuracy to over 99%.
- Led a team of 3 engineers in developing an SRGAN+CNN image classification method, improving throughput of SEM review by 4X.

EDUCATION

M.S. Optoelectronic Engineering, *San Jose State University* 2020
M.S. Materials, *University of California, Santa Barbara* 2015
M.S. Materials Engineering, *Cum Laude,* *California Polytechnic State University* 2012
B.S. Materials Engineering, *Honors with Distinction,* *California Polytechnic State University* 2011

PROJECT WORK

OutcoGPT Technical Interview Chatbot | AI Machine Learning Engineer | github 2023

An interview chatbot based on ChatGPT that uses the technical interview expertise at Outco.io.

- Developed a mock code interview chatbot using ChatGPT with user speech input and generated audio responses, grading user code inputs in real-time using a Gradio interface.