Pranav Arora
pranav2vis@gmail.com · +65 85157891 · linkedin.com/in/pranavarora63 · pranavarora.vercel.app

Experience

Senior ML Engineer - AI Engineering Team **Aug 2024 – Present**
Hewlett Packard Enterprise · Singapore

Text-to-SQL AI Platform

- Architected an end-to-end natural language query platform enabling 200+ non-technical users across supply chain, logistics, and operations to query databases without SQL, reducing analyst dependency by 40%
- Implemented LangChain SQL agent with GPT-4o using few-shot prompting and schema-aware generation across 75+ tables, achieving 85% query accuracy
- Built Redis caching for frequent aggregations (30% cache hit rate) and read-replica routing to protect production databases
- Deployed on HPE Ezmeral Kubernetes cluster with FastAPI serving layer; integrated Keycloak SSO and row-level security enforcement
- Managed Azure AI Foundry model deployments and configured RBAC for team-scoped LLM access across multiple Azure OpenAI endpoints

Impact: 2,000+ queries/week · 7 business units · 40 analyst hours saved/week · average query time reduced from 10 min → 1 min

K8s Watcher - Agentic Kubernetes Monitoring & Remediation

- Designed and deployed a production agentic AI system (FastAPI + React) that autonomously monitors Kubernetes cluster health, diagnoses incidents using GPT-4o, and surfaces remediation recommendations in real time
- System integrates with live cluster state (pod logs, events, resource metrics) to provide context-aware root cause analysis across namespaces
- Backed by PostgreSQL for incident history and trend analysis; deployed on HPE Ezmeral with Keycloak-authenticated access

Impact: 50+ incidents diagnosed/week · MTTR reduced by 70% · replaced manual kubectl triage for on-call engineers

Document Planning Hub - Multi-Agent Validation System

- Built a multi-agent orchestration system using LangGraph and Azure OpenAI to automate validation of enterprise documents (PDF, PPTX, Excel, DOCX) - checking broken links, font alignment, structural compliance, and formatting rules
- Designed agent workflow as a directed graph: dispatcher agent → specialist agents per document type → aggregator agent producing structured validation reports
- Deployed as a FastAPI microservice integrated into the OneAI platform with Keycloak auth and Kubernetes-native scaling

Impact: 5,000+ users across Digital Planning and Marketing teams · 80% reduction in document errors · 10 hours saved/week for content teams

Real-Time Address Correction

- Deployed a Python-based LLM solution integrating Google Maps API and Azure OpenAI to validate and correct delivery addresses in real time during order processing
- Hosted on Azure AKS with CI/CD pipelines on Azure DevOps; designed for high-throughput order ingestion with sub-second correction latency

Impact: 35% reduction in supply chain fallouts · 50% fewer manual interventions

OneAI Platform Engineering Standards

- Led AI engineering standards across 8 teams: standardized CI/CD templates (GitHub Actions + Azure DevOps), MLOps patterns, Docker build conventions, and code review frameworks
- Built AzureOpenAI oncall AI agent for suggestive autonomous production fixes; established bi-weekly tech reviews across 20+ AI applications

Impact: Onboarding time reduced 40% · deployment failures down 60% · time-to-production from 6 weeks → 2 weeks

Data Scientist - Smart Manufacturing & AI Team **Jan 2022 – Aug 2024**
Micron Technology · Singapore

- Optimised semiconductor wafer scheduling using reinforcement learning agents (PPO via Ray RLlib) across a 70-machine distributed training cluster, owning design through production deployment
- Deployed agent recommendation pipeline updating production UI every 2 hours; refactored NumPy-based simulation code reducing runtime by 60%
- Fine-tuned an LLM chatbot on 10K+ internal documents (manuals, tickets, product data) for manufacturing knowledge access, achieving BLEU score of 0.82 and 80% first-contact resolution on support queries
- Standardised master data across planning and operations applications, reducing data input and maintenance by 80% and cutting time spent searching for metrics by 50%

Impact: 0.5% wafer production increase → $10M annual revenue · 15% reduction in queuing time

Data Scientist - Data Analytics Team **Aug 2020 – Jan 2022**
Dentsu International · Singapore

- Developed a customer purchase propensity model (85% validation accuracy) deployed to production and integrated into live marketing campaigns
- Deployed a modern data catalog tool to Azure AKS using Terraform, ingesting 10,000+ datasets from Azure Synapse for enterprise data governance
- Built ROAS prediction models and reporting dashboards reducing post-campaign analysis time by 50% and media planning costs by 20%

Education

Master of IT in Business (Major: Artificial Intelligence) - Singapore Management University Jan 2019 – May 2020
Bachelor of Computer Science - University of Petroleum and Energy Studies, India Aug 2014 – Apr 2018

Skills

Azure (primary cloud): Azure AI Foundry, Azure OpenAI Service, Azure ML Studio, Azure Cognitive Services (Speech, Vision), Azure Communication Services, AKS, Azure DevOps, Azure Synapse, Azure Container Apps
Agentic AI & LLMs: LangGraph, LangChain, AutoGen, Multi-Agent Orchestration, RAG systems, GPT-4o, Azure OpenAI, Hugging Face Transformers, Prompt Engineering, LangSmith
ML & Modeling: Reinforcement Learning (PPO, Ray RLlib), TensorFlow, Scikit-Learn, XGBoost, Keras, Statistical Modeling, Fine-tuning, SHAP/LIME
Cloud & Infra: Kubernetes, Docker, Istio, AWS (SageMaker, Bedrock, EKS), GCP (Vertex AI, BigQuery), HashiCorp Vault, Keycloak, Istio Service Mesh
MLOps & Deployment: MLflow, KServe, BentoML, TorchServe, Prometheus, Grafana, GitHub Actions, Azure DevOps CI/CD, A/B Testing, Model Versioning
Data & Vector Stores: PostgreSQL (pgvector), Pinecone, ChromaDB, Snowflake, MySQL, Redis, Apache Airflow, Prefect, Kubeflow Pipelines
Languages & Frameworks: Python, SQL, Bash, JavaScript/ReactJS, FastAPI, Flask, REST APIs, Microservices, Event-Driven Architecture
Visualization & Other: Tableau, PowerBI, Plotly, Streamlit, OpenCV, YOLO, OCR (Tesseract, PaddleOCR), N8N, Microsoft Copilot Studio, spaCy, NLTK

Certifications & Publications

Harvard Business School - Business Analytics · *Innovations in Software Defined Networking and Network Virtualizations* (Publication)