Jian Wang
https://www.linkedin.com/in/jianwang92/
jianwang.data@gmail.com
Menlo Park, CA

Experience

DoorDash, Senior Machine Learning Engineer Sep 2024 – present

- **LLM Product Matching**: Finetuned and deployed GPT-4o mini for matching merchant products to DoorDash catalog, improving link rate by 18% and unlocking $174M in annual ad revenue
- **Real-Time Vector Search**: Built end-to-end retrieval pipeline with GPU-served CLIP embeddings and Qdrant vector database, powering onboarding of 1.3M products weekly at 200 QPS peak
- **Open-Source LLM Finetuning**: Finetuned Llama and Qwen using LoRA and quantization to replace commercial APIs, reducing cost by 87% and improving service stability
- **Agentic VLM Matching**: Prompt engineered GPT-5.2 VLM with agentic web search to identify missing information and process the hardest matching tasks. Statistically outperformed human annotators at 63% win rate, with potential to save $500k in annual labeling costs
- **AI Reading Workshop**: Hosting company-wide workshop, growing attendance 3x to 150 engineers per session. Invited internal and external speakers on LLM applications and production AI

GoPuff, Senior Data Scientist, Search and Discovery Jan 2022 – Jul 2024

- **Query Understanding**: Developed BERT-based NLU for intent classification and entity extraction. Built LLM annotation pipelines to scale training data, deploying to 5 services and driving 0.90% revenue lift
- **Product Retrieval**: Architected hybrid retrieval with BM25 and ANN-indexed embeddings. Integrated intent classification and entity tagging, driving 0.33% search conversion improvement
- **Search Ranking**: Built personalized ranking system, evolving from XGBoost to Wide & Deep neural architecture. Owned end-to-end ML lifecycle from feature engineering to production deployment. Drove $600k savings, 1.03% conversion lift, and 0.47% margin gain
- **Multi-Objective Optimization**: Designed optimization layer to jointly balance conversion, margin, and exploration objectives. Built simulation framework for tuning. Lifted margin by 0.31%
- **Product Similarity**: Designed contrastive fine-tuning pipeline for transformer-based product embeddings. Powered real-time out-of-stock substitution, reducing cancellations by 1.19%

LivePerson, Data Scientist II, Conversational AI Nov 2020 – Jan 2022

- **Multilingual NLU**: Built intent classification and entity tagging models in 8 languages. Deployed in contact center chatbots across 4 industry verticals, driving 5.2% intent resolution improvement
- **Few-Shot Entity Tagging**: Developed contrastive learning approach for entity tagging with as few as 5 examples per type. Reduced customer annotation requirements by 50%

Amazon, Software Development Engineer, Machine Learning Apr 2020 – Nov 2020

- **ML Training Platform**: Built serverless training infrastructure on Lambda, DynamoDB, and S3 for Alexa organization. Enabled scientists to run model training workflows with self-service provisioning
- **Privacy-First Access Controls**: Implemented dataset-level permission system enforcing data governance policies, enabling training on sensitive user data with full audit compliance

LivePerson, Data Scientist, Conversational AI Nov 2018 – Apr 2020

- **Text Classification**: Developed deep learning package for intent detection and text classification (PyTorch, BERT finetuning). Achieved <5 ms inference across architectures
- **Anomaly Detection**: Implemented confidence calibration layer on intent classifiers to flag low-confidence predictions, reducing erroneous bot responses. Achieved 2% accuracy gain over prior rule-based approach

Skills

LLMs & GenAI: Fine-tuning (LoRA, QLoRA), RAG, Prompt Engineering, Agentic Systems, VLMs

ML Frameworks: Python, PyTorch, HuggingFace, TensorFlow, XGBoost, LightGBM

Production ML: FAISS, Qdrant, Milvus, CLIP, W&B, Ray, Triton, Arize

Infrastructure: AWS, Databricks, Snowflake, Docker, Kubernetes, SQL, PySpark

Awards

Gold Medal, Chinese Physics Olympiad (2011): 51 winners in China among thousands of competitors

Publications and Patents

Domain adaptation of AI NLP encoders with knowledge distillation. Kristen Howell, **Jian Wang**, Matthew Dunn, Joseph Bradley. *United States Patent US-11568141-B2, 2023.*

Domain-Specific Knowledge Distillation Yields Smaller and Better Models for Conversational Commerce. Kristen Howell, **Jian Wang**, Akshay Hazare, Joseph Bradley, Chris Brew, Xi Chen, Matthew Dunn, Beth Ann Hockey, Andrew Maurer, Dominic Widdows. *e-Commerce and NLP (ECNLP), 2022.*

Think Visually: Question Answering through Virtual Imagery. Ankit Goyal, **Jian Wang**, Jia Deng. *Association for Computational Linguistics (ACL), 2018.*

Premise Selection for Theorem Proving by Deep Graph Embedding. Mingzhe Wang, Yihe Tang, **Jian Wang**, Jia Deng. *Neural Information Processing Systems (NeurIPS), 2017.*

Education

University of Michigan Ann Arbor, MI
Master of Science in Computer Science Sep 2015 – Aug 2018

Peking University Beijing, China
Bachelor of Science in Physics Sep 2011 – Jun 2015