Zhijin (Daniel) Fang

310-923-3337 | daniel.zhijin.fang@gmail.com | linkedin.com/in/daniel-fang-85748a205/ | github.com/danielfang001

Education

University of California, Los Angeles — Los Angeles, CA
Bachelor of Science in Data Theory, Minor in Data Science Engineering ***GPA: 3.92/4.0*** — *Sep. 2021 – Dec. 2025*

Experience

Machine Learning Engineer Intern — July 2024 – Sep. 2024
Insilico Medicine — *Shanghai, China*

- Conducted research on state-of-the-art **RAG** algorithms, evaluated on PubMed research paper retrieval and question answering capabilities by assessing recall rate and precision using the Ragas framework
- Optimized paper parsing by enhancing **OCR** and **chunking**. Deployed an offline **Elasticsearch** database for efficient sparse & dense vector searches across multimodal corpus of **50k+** biomedical papers. Collaborated with open-source on unstructured data handling through indexing and sharding techniques
- Developed a **agentic RAG** pipeline using **Python**, **Llama 3.1**, **Docker**, **REST APIs (Flask)** and **Langchain** to mitigate hallucinations on domain-specific named entities, resulting in a **20% increase** in recall and precision

Co-founder & Lead AI/ML Engineer — July 2024 – Present
Ult.ai – the first AI-empowered natural language search engine for social discovery — *Los Angeles, CA*

- Led research and development of Ult's core human-profile search, capable of interpreting natural language queries, understanding user intent, utilizing high-dimensional **hybrid embeddings** and **CoT** for optimized search results
- Deployed **production-scale** search pipeline with **Python**, **OpenAI APIs (structural output)** and **Flask** to perform query preprocessing, talking to **Milvus** database in natural language and results reranking
- Developed non-blocking **asynchronous** search task scheduling system via **Golang**, **Echo**, **RabbitMQ** and **MySQL**, decoupled search process from the main user flow to maintain seamless user experience and optimized user sign-up session to use third-party email API via **Redis with TTL**

R&D Data Analyst Intern — June 2023 – Sep. 2023
TerraCycle — *Trenton, NJ*

- Led data-driven research on supply chain for a high-impact container recycling project with Walmart. Developed a **analytical interface** using **Python**, **Pandas**, **Scikit-learn** and **matplotlib** to visualize financial data

Business Consultant Intern — Dec. 2022 – Mar. 2023
Ernst & Young — *Shanghai, China*

- Collaborated with AstraZeneca to analyze **500+** pharmaceutical companies in China. Using web-scraping (**Selenium**), rating algorithms and **Tableau** to perform **risk assessments** using financial data and credit records. Authored 50+ reports on target companies' backgrounds and qualifications for partnership consideration

Projects

LLM Inference Optimization: BUZZ KV Cache | *Pytorch, Llama, Optimization* — July 2024 – Present
arXiv:2410.23079 **(in review)**

- Developed a **new KV caching algorithm** that optimizes LLM cache usage during real-time inference
- Reduced GPU memory usage by over 2.5×, achieving log(n) time complexity and exceeding **over 99% accuracy** in long-text summarization. Achieved a **7.69%** performance improvement over naive models in multi-document question answering, **surpassing state-of-the-art** methods under the same cache memory constraints

AI-generated Text Detection with DeBERTaV3 | *Python, Keras, WanDB, BERT* — May 2018 – May 2020

- Finetuned a **task-specific DeBERTaV3** classifier using KerasNLP. Used tokenizer and masked embedding for preprocessing and leveraged adaptive learning rate for downstream training. Validation accuracy reached **99.97%**

Technical Skills

Languages: Python, C++, SQL, Golang, JavaScript, R, DBML, React.js
Databases: MySQL, PostgreSQL, Neo4j, MongoDB, Milvus, Elasticsearch, Redis
Libraries: Pandas, NumPy, matplotlib, PyTorch, Tensorflow, huggingface, sklearn, selenium, request, Flask, PySpark
Other: Shell, Kafka, Docker, Git, AWS EC2&S3, Apache Spark/Flink CDC/Airflow, Hadoop, Databricks, Tableau, Linux, Prompt Engineering, LLMs, Parellel Computing, Database Normalization, ETL Automation, Hypothesis Testing