Runpeng (Benson) Jian

runpengj@gmail.com | +1 (650) 516-5771 | linkedin.com/in/runpengjian/ | github.com/RunpengJ | San Diego, CA

EDUCATION

University of California, San Diego — *La Jolla, CA*
M.S. Computer Science, GPA: 3.99 — Sep 2024 – Dec 2025
B.S. Computer Science, GPA: 3.85 — Sep 2022 – Aug 2024
Honors/Scholarships: Future Leaders in STEM Scholarship, Qualcomm Alumni Scholarship

EXPERIENCE

ChatDocs — Jun 2025 – Present
Machine Learning Engineer (Team Project) | FastAPI, AWS CDK, OpenSearch, Bedrock, Lambda — *La Jolla, CA*

- Designed and shipped a production-grade RAG platform on AWS (S3, Lambda, OpenSearch, Bedrock) with fully automated ingestion and semantic search; scaled distributed processing to 1M+ documents with resilient batching, retries, and DLQs.
- Built modular services for embeddings, retrieval, and prompt orchestration; integrated Amazon Titan embeddings with OpenSearch k-NN to enable low-latency semantic search and classification workflows.
- Established CI/CD and test automation (unit/integration/contract) with Docker and AWS CDK, achieving 90%+ coverage; added structured logging, metrics, and alerts for deployment safety and runtime observability.
- Built an LLM evaluation harness (correctness, grounding, latency, cost) with CI gates to prevent regressions; added input/output validation and safety filters for prompt and response control.

Outlier AI — May 2024 – Mar 2025
AI Trainer | RLHF, Prompting, LLM Evaluation — *Remote*

- Designed and implemented structured prompt engineering workflows and reinforcement learning from human feedback (RLHF) methodologies, boosting code generation accuracy by 17% for complex classification challenges.
- Established automated code-review and quality-assurance processes, auditing 500+ AI-generated code samples/week and resolving performance bottlenecks and style violations.
- Developed targeted prompt strategies and standardized output-rewriting guidelines, raising task success rates by 25% and ensuring alignment with PyTorch and TensorFlow best practices.

Wang Lab — Jun 2023 – Sep 2023
Research Engineer | PyTorch, Transformer, Diffusion Models — *UC San Diego, La Jolla, CA*

- Evaluated 15+ neural network architectures using PyTorch for image processing, contrasting performance metrics like PSNR and SSIM, pinpointing optimal methods for visual reconstruction tasks.
- Built end-to-end training and evaluation workflows for diffusion models, processing large-scale image datasets (10K+ samples) and conducting ablation studies to optimize model performance for inverse problem applications.
- Synthesized findings from 50+ research papers on diffusion models, developed performance comparison framework, and presented technical insights at NCUR and UCSD Summer Research Conference.

Ujima Lab — Sep 2022 – Jun 2023
Machine Learning Engineer | Python, NLP, Data Mining — *UC San Diego, La Jolla, CA*

- Leveraged NLP techniques including sentiment analysis, topic modeling, and text classification to analyze 75K+ Reddit comments, uncovering 8 distinct privacy concern patterns through unsupervised clustering algorithms using scikit-learn.
- Built automated data collection and processing pipeline using Python, Reddit API, pandas, and NLTK to gather user experiences across 15+ gaming platforms, implementing scalable text processing workflows for dataset creation.
- Co-authored paper accepted at WIPS 2023, contributing ML-driven visualizations and achieving 85% accuracy in automated privacy concern classification.

PROJECTS

Image Synthesis with Diffusion Models | *Pytorch, Diffuser, LoRA/DreamBooth*

- Trained and fine-tuned Stable Diffusion for character generation on 50K+ samples; optimized with DreamBooth and distributed GPU training for a 15% improvement in output diversity/clarity.
- Implemented batch workflows and model optimizations to reduce training time by 30% while maintaining semantic alignment and quality across architectures.

Recommendation System for Local Business Analytics | *Gradient Boosting, Sentiment Analysis, Machine Learning*

- Achieved 20% accuracy improvement and 15% RMSE reduction in rating predictions by combining collaborative filtering with sentiment analysis over 100K+ reviews and business metadata.
- Built scalable feature-engineering and model-training pipelines for offline batch inference; implemented distributed data workflows for similarity computation and periodic retraining across dozens of business categories.

SKILLS

- **ML/AI:** Retrieval-Augmented Generation (RAG), Machine Learning, Deep Learning, NLP, RLHF, Recommender Systems, Prompt Engineering, LLM Evaluation
- **Cloud/DevOps & Systems:** AWS (Lambda, S3, OpenSearch, Bedrock, EC2/Spot), Docker, GitHub Actions, IaC (AWS CDK), Linux/UNIX, Distributed Systems, Model Serving, CI/CD
- **Frameworks:** PyTorch, TensorFlow, Hugging Face, FAISS, Diffusers, FastAPI
- **Languages:** Python, C++, Java, SQL, JavaScript/TypeScript

LEADERSHIP & AFFILIATIONS

Codepath | Community Member — Jun 2025 – Aug 2025
UC LEADS Program | Scholar — Jun 2023 – Jun 2024
Chinese Engineering Society | Board Member — Sep 2022 – Jun 2023