

Daozheng Chen

AI/ML & Full Stack Engineer | LLM & MLOps Expert

+1 408 786 2961 | dzh.chen.ai@gmail.com | [LinkedIn](#) | San Jose, CA

SUMMARY

Senior AI/ML and Full-Stack Engineer with 16+ years of experience designing and delivering large-scale machine learning platforms, LLM-powered search systems, and recommendation infrastructures at Meta and Yahoo. Proven expertise in architecting end-to-end AI solutions, including data pipelines, model training, inference services, and product-grade APIs for Ads, Search, and Recommendation systems serving hundreds of millions of users. Skilled in deploying RAG pipelines, transformer-based NLP models, real-time inference platforms, and scalable MLOps workflows with a strong focus on performance, reliability, and cost efficiency. Experienced in leading full-stack development, building internal ML tooling, and collaborating cross-functionally to productionize research models while ensuring security, compliance, and operational excellence.

EXPERIENCE

Research Scientist (LLM & Search Systems)

Yahoo | San Jose, CA

06/2025 – Present

- Built LLM-augmented search and question-answering systems using PyTorch, Hugging Face Transformers, LangChain, Milvus/FAISS, and FastAPI, improving semantic relevance (NDCG over 14%) across Yahoo Search and content platforms.
- Designed and deployed RAG pipelines combining vector search, BM25 hybrid retrieval, and transformer re-ranking, reducing hallucination rate by 30% and answer latency by 35%.
- Developed GPU-optimized inference services using Kubernetes, Triton/ONNX Runtime, and NVIDIA A10/T4 GPUs, supporting 20K+ QPS with P99 latency of less than 120ms.
- Implemented LLM evaluation and A/B experimentation frameworks using Python, BigQuery, and Prometheus, enabling data-driven rollout decisions across millions of daily users.
- Partnered with product and infra teams to productionize research models with CI/CD (GitHub Actions), observability (Grafana), and cost-aware scaling, reducing inference cost per request by 25%.
- Orchestrated end-to-end machine learning pipelines using Python, AWS SageMaker, and Bedrock to process real-time user interaction and content signals, improving high-risk and low-quality content detection accuracy by 15% and pipeline efficiency by 25%.

Machine Learning Engineer & Full Stack Engineer

Meta | Sunnyvale, CA

02/2016 - 06/2024

- Architected large-scale ML ranking and recommendation systems for Ads, Search, and Feed, processing billions of daily events using PyTorch, TensorFlow, Kafka, and distributed feature stores.
- Enhanced AI-driven monitoring and evaluation workflows by integrating GCP Vertex AI and Azure Cognitive Services for advanced NLP and computer vision research, achieving 20% higher model accuracy in offline and online benchmarks.
- Designed and evaluated predictive and NLP models using Scikit-learn, Hugging Face Transformers, and Snowflake, improving triage and relevance classification quality across large-scale content and QA systems.

- Built end-to-end ML pipelines with Airflow, Spark, BigQuery, and custom feature stores, accelerating model iteration velocity by 40%.
- Developed deep learning models (DNNs, Transformers) for ad relevance and engagement prediction, delivering +8–12% CTR lift in global A/B experiments.
- Streamlined large-scale analytics using GCP BigQuery and cloud-native ML workflows, improving forecast and trend-detection accuracy by 12% across research datasets.
- Led development of multiple GenAI and LLM prototypes using TensorFlow and Hugging Face, achieving up to 90% model accuracy while reducing inference latency by 25ms.
- Implemented low-latency real-time inference services using Python, Go, FastAPI, ONNX Runtime, and gRPC, achieving P99 latency under 20ms at production scale.
- Designed online experimentation and model rollout frameworks enabling safe deployment to hundreds of millions of users with automatic rollback and metric-based gating.
- Built internal full-stack tools using React, TypeScript, GraphQL, and Node.js/Express, improving ML debugging and monitoring workflows for 100+ engineers.
- Led architecture reviews and mentored senior engineers on ML system design, reliability, and scalability, contributing to long-term ML platform strategy.
- Designed and evaluated predictive and NLP models using Scikit-learn, Hugging Face Transformers, and Snowflake, improving triage and relevance classification quality across large-scale content and QA systems.
- Integrated deep learning models with internal dashboards and tooling built using React, TypeScript, Next.js, and Redux, enabling engineers and product teams to debug, monitor, and analyze ML systems more effectively.
- Implemented computer vision pipelines using PyTorch, OpenCV, and transformer-based models, reducing data quality and classification errors by 50% in production ML workflows.
- Built and optimized end-to-end ML pipelines using Kubeflow, Azure DevOps CI/CD-style workflows, and internal orchestration systems, improving model reliability while reducing operational costs by 20%.

Senior Software Engineer

Yahoo | San Jose, CA

04/2013 - 02/2016

- Built content processing and personalization pipelines using Python, Java, Hadoop, and Spark, powering Yahoo Homepage, News Feed, and Yahoo Recommends for tens of millions of users.
- Built and maintained full-stack web applications to process, analyze, and visualize large-scale user and content data using Python (Django/Flask) for backend services and ETL pipelines, improving internal data insights and decision-making by 25%.
- Developed and consumed RESTful APIs using Node.js/Express, applying security best practices and automated CI/CD pipelines with GitHub Actions, while integrating Kafka-based real-time data streams processing 300K+ events with 99.8% reliability.
- Owned production reliability and time-sensitive releases, executing content updates and database migrations with 99.9% uptime, leveraging AWS-based infrastructure and Docker for consistent, repeatable deployments.
- Designed, implemented, and debugged scalable backend services using Go, integrating securely with APIs and data stores, and deploying via Kubernetes-managed microservices, reducing production defects by 80%.
- Developed NLP models for news relevance scoring, topic classification, and low-quality content detection using TF-IDF, logistic regression, and early deep learning models, increasing personalization precision by 18%.
- Designed and trained CNN-based computer vision models using Caffe/TensorFlow for smart image cropping and nudity detection, deployed in real-time content moderation systems.
- Implemented financial news relevance ranking systems that improved user engagement and session duration by 10%.

- Built scalable ETL pipelines and real-time scoring services integrated with downstream personalization and recommendation platforms.

Research Intern

Siemens Corporate Research | Princeton, NJ

05/2012 - 08/2012

- Conducted applied ML research using Python, MATLAB, and probabilistic models, improving pattern recognition accuracy by 15% on industrial datasets.
- Prototyped and validated algorithms on large-scale real-world data for predictive analytics use cases.

Research Intern

Toyota Technological Institute at Chicago | Chicago, IL

08/2011 - 11/2011

- Built computer vision and statistical learning models using Python and MATLAB for visual recognition and feature extraction tasks.
- Contributed to experimental research evaluating model robustness and generalization on academic datasets.

Research Intern

MIT Computer Science and Artificial Intelligence Laboratory | Cambridge, MA

09/2010 - 03/2011

- Conducted advanced ML research in collaboration with PhD researchers, implementing prototype ML systems using Python and C++.
- Produced internal research artifacts supporting ongoing AI systems and learning theory research.

EDUCATION

Doctor of Philosophy - PhD, Computer Science

University of Maryland | 2009 – 2013

Master of Science in Computer Science

University of Maryland | 2007 – 2009

Bachelor of Science in Computer Science

University of Maryland | 2003 – 2007

SKILLS

- **Machin Learning Technologies:** PyTorch, TensorFlow, NLP, Hugging Face Transformers, Scikit-learn, GenAI, LangChain, Model Serving, RAG, AutoML, ONNX Runtime, Computer Vision, Statistical Modelling
- **Programming Languages & Frameworks:** Node.js, TypeScript, React, Redux, Python, Go, Express, FastAPI, Next.js, NestJs, Angular, RESTful APIs, GraphQL, Java, SpringBoot
- **Cloud Technologies:** AWS, GCP, Azure, Serverless Architectures, API Gateways, Cloud Mornitoring (Prometheus, Grafana), Infrastructure as Code

- **DevOps & Deployment:** Docker, Kubernetes, Helm, GitHub Actions, CI/CD, Containerization, Orchestration, Serverless Architectures, Infrastructure as Code, Monitoring & Logging (ELK Stack), Prometheus, Grafana
- **Data Engineering & Databases:** Data Warehousing, ETL, Data Pipelines, Data Modeling, Schema Design, PostgreSQL, MongoDB, Redis, GCP BigQuery, Weaviate, Milvus, Data Encryption
- **API & Backend Technologies:** RESTful APIs, GraphQL, Node.js (Express), Go (Golang), Python (FastAPI), API Design, API Gateways, Authentication, Authorization, Performance Tuning, Concurrency Patterns
- **Frontend Technologies:** React, Redux, TypeScript, Next.js, Angular, Material UI, Tailwind CSS, Context API, Interactive Frontend Components, Single Page Applications (SPAs)
- **Security & Compliance:** HIPAA Compliance, Data Encryption, Authentication, Authorization, AWS Security Services, GCP Security Features, Azure Security Center, Vulnerability Scanning, Penetration Testing, Security Audits

Key Projects & Achievements

- **LLM-Augmented Search, Chat, and Knowledge Retrieval Platform:** Designed and deployed a production-grade LLM-powered search and conversational QA system using PyTorch, Hugging Face Transformers, LangChain, prompt engineering, and RAG pipelines backed by FAISS, Milvus, Weaviate, and Pinecone. Combined BM25 hybrid retrieval, transformer re-ranking, and context grounding to reduce hallucinations by 30% and improve semantic relevance (NDCG +14%). Built Python/FastAPI and Node.js services with React + Redux / Angular frontends, deployed on Kubernetes, serving millions of queries with sub-150ms P99 latency.
- **Large-Scale Personalized Recommendation & Ranking Systems:** Architected recommender and ranking systems using deep learning (DNNs, Transformers), Scikit-learn, NLP models, and GenAI-assisted feature enrichment. Implemented Go and Python microservices, RESTful APIs, GraphQL, and Kafka-based event pipelines for real-time user behavior ingestion. Stored features and embeddings in PostgreSQL, MongoDB, Redis, and vector databases, achieving 8–12% CTR lift and 21% efficiency gains in large-scale simulations and A/B tests.
- **Computer Vision–Driven Content & Inventory Intelligence Pipelines:** Built computer vision systems using PyTorch, TensorFlow, OpenCV, CNNs, and transformer-based vision models for image classification, quality detection, smart cropping, and object recognition. Integrated CV outputs into real-time ML pipelines, AWS SageMaker deployments, and Dockerized microservices, reducing classification and data quality errors by ~50% across large content and inventory datasets.
- **Real-Time ML Inference, Model Serving & MLOps Infrastructure:** Developed GPU-optimized inference services using Kubernetes, Triton Inference Server, ONNX Runtime, and AWS Bedrock, supporting 20K+ QPS. Implemented CI/CD with GitHub Actions, Helm-based deployments, autoscaling, secrets management, and cloud monitoring with Prometheus, Grafana, and ELK Stack, cutting inference cost per request by 25% while maintaining strict latency SLAs.
- **End-to-End ML Pipelines, Analytics & Experimentation Frameworks:** Built scalable ML pipelines using Airflow, Spark, BigQuery, Snowflake, and cloud-native orchestration, supporting training, evaluation, and deployment of NLP, CV, and ranking models. Designed A/B experimentation frameworks, automated rollout gates, and rollback mechanisms used across hundreds of millions of users, improving iteration velocity by 40%.
- **Event-Driven Microservices & Data Platforms:** Designed a high-throughput event-driven architecture using Go (concurrency-optimized services), Kafka consumers/producers, MongoDB, PostgreSQL, and REST APIs, processing 2M+ events/day with 99.99% availability. Deployed systems on Kubernetes with Helm, enabling fault tolerance, schema evolution, and backpressure handling at scale.
- **Security, Compliance & Production Readiness:** Implemented HIPAA-aware data handling, authentication/authorization, data encryption, and security best practices across AWS, GCP, and Azure environments. Integrated vulnerability scanning, audit logging, and secure API gateways into ML and data platforms used in regulated and high-risk domains.
- **Internal ML Tooling & Full-Stack Platforms:** Built internal platforms and dashboards using React, TypeScript, Next.js, Angular, Redux, Material UI, Tailwind CSS, and Node.js/Express backends. Enabled engineers and product teams to debug, monitor, and analyze ML systems, model performance, and data quality, significantly improving operational visibility and developer productivity.