Shane (Xingeng) Zhang

Piscataway, NJ 08854 | zhangxingeng970221@gmail.com | 732-491-6378
www.shanechang.com | github.com/zhangxingeng | https://www.linkedin.com/in/zhangxingeng

SUMMARY

Full-Stack AI Engineer with 5+ years building production LLM-powered systems. Specializing in **agentic AI architectures** (tool use, planning, memory), **RAG pipelines**, and **distributed workflow orchestration**. Built systems serving 2,000+ users with 1M+ daily API calls.

SKILLS

GenAI: Agentic AI Systems (Tool Use, Planning, Memory), Prompt Engineering (Chain-of-Thought, Few-Shot), RAG Pipelines, Multi-Agent Orchestration, LLM Evaluation, Structured Output Generation
Languages: Python, TypeScript, JavaScript, SQL, C++ **Frameworks:** FastAPI, Litestar, SvelteKit, React, Node.js, Temporal.io
ML/AI: PyTorch (Distributed), Transformers, HuggingFace, LangChain, Pydantic-AI, LiteLLM **Data:** PostgreSQL, Redis, Vector Databases (Chroma, pgvector), Elasticsearch, Neo4j
DevOps: Docker, Kubernetes, OpenTelemetry, GitHub Actions, AWS (S3, SageMaker, Bedrock), Azure

WORK EXPERIENCE

Empath Legal (empathlegal.com) Piscataway, NJ
Founding Engineer (Full-Stack AI) *Dec 2024 - Present*

- Built end-to-end **agentic AI system** with **tool use, planning, and memory**: **Python/Litestar** backend, **SvelteKit/TypeScript** frontend, **PostgreSQL**, orchestrating 78+ GenAI pipelines for juror research.
- Designed **multi-agent workflow orchestration** with **Temporal.io**: fault-tolerant pipelines, human-in-the-loop feedback, **Chain-of-Thought prompting**, step-level pause/resume.
- Architected **multi-provider LLM layer** with **LiteLLM**/**Pydantic-AI**: seamless switching between **GPT-4/5**, **Claude 3/Opus**, **Gemini**, **Mistral**, **Llama 3** with unified structured output and tool calling.
- Implemented **type-safe architecture**: OpenAPI-generated TypeScript schemas, Pydantic models driving API contracts; real-time **SSE** streaming with Redis pub/sub.
- Engineered production infrastructure: **Docker Compose** (10+ services), **OpenTelemetry** observability, OAuth 2.0 (Google, Microsoft, Apple), **353 automated tests**.

Citigroup Inc. Rutherford, NJ
AI / ML Engineer *Aug 2024 - Dec 2024*

- Built **Agentic RAG system** with Knowledge Graphs and **vector databases** (Elasticsearch, pgvector) for compliance analysis, improving F1 **0.32→0.71**, reducing hallucinations by **30%**.
- Designed **FastAPI** backend: **1M+ API calls/day**, reducing peer review from **4 hours to minutes** per document; zero downtime over 120 days.
- Built data pipeline reducing document processing from 10s to **<1 second**, serving **2,000+ users**; deployed on **AWS** (SageMaker, Bedrock) with Terraform.

Robert Wood Johnson University Hospital New Brunswick, NJ
AI / ML Engineer *Jan 2023 - Aug 2024*

- Built **RAG system** with **Chroma** automating summarization of **5,000+ research papers**, accelerating literature review for neuroscience researchers.
- Trained custom **multi-modal transformer** for categorizing neuroscience data (images + time series), contributing to published research.
- Collaborated with **15 researchers**, integrating **200+ feedback points**; built **React** frontend reducing onboarding time by **50%**.

Rutgers University - Research Lab New Brunswick, NJ
Graduate Research Assistant *Sep 2021 - Dec 2022*

- Researched **transformer-based models** for 3D point cloud reconstruction; implemented **PyTorch Distributed** training across multi-GPU clusters, reducing training time by **40%**.

Fiskkit Inc. San Francisco, CA
Machine Learning Engineer *Jan 2020 - July 2021*

- Integrated **NLP features** into **Node.js** backend with **PyTorch** (CUDA), enabling real-time text generation and summarization, cutting response times by **70%**.
- Built data pipelines with **PySpark** and **PyTorch Distributed**, improving model training results by **20%** over baseline.

EDUCATION

Rutgers University, New Brunswick, NJ Dec 2020 - Dec 2022
M.S. Computer Science (Machine Learning) GPA: **3.71**
Rutgers University, New Brunswick, NJ Sep 2017 - Sep 2020
B.S. Computer Science

ADDITIONAL

Certification: AWS Certified Machine Learning - Specialty (MLS-C01) Jun 2024
Portfolio: shanechang.com/portfolio – Technical blog with AI-generated content (Hugo, Tailwind)
Teaching: TA at Rutgers, mentored **400+ students** in CS/ML; experienced onboarding and leading engineers
Languages: English (fluent), Mandarin (native)