

Barry (Yuanshen) Wang

Software Engineer — AI

barryep2025@gmail.com | barryep.com/about | Irvine, CA

Software engineer with 3 years of hands-on experience shipping production generative-AI, LLM application, and AI agent systems. Strong Python plus Java; AI agent orchestration, multi-agent collaboration, RAG architectures, vector databases, prompt and context engineering, and LLM evaluation. Builds REST APIs, microservices, and event-driven architectures on AWS.

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, SQL, C/C++, Solidity

AI / Machine Learning & Agents: LLM application development, AI agent orchestration, agentic workflows, multi-agent collaboration patterns, prompt engineering, context engineering, memory management, Retrieval Augmented Generation (RAG), embeddings and semantic search, AI model evaluation and benchmarking, NLP, LangChain, Claude Agent SDK, Model Context Protocol (MCP), Anthropic and OpenAI platforms

Python Ecosystem & Data: Pandas, NumPy, LangChain, data processing and analytics, ETL and data pipelines, structured and unstructured datasets, statistical analysis, data visualization, JSON/XML processing

Databases: PostgreSQL, SQL and NoSQL databases (AWS DynamoDB), Vector Databases (Zilliz / Milvus), Elasticsearch

APIs & Backend: REST APIs, microservices, event-driven architecture, FastAPI, Java Spring Boot, NestJS, Node.js / Express, authentication and authorization

Cloud & DevOps: AWS (Lambda, Fargate, SQS, API Gateway, RDS/Aurora, S3, CloudFront, EventBridge), Docker, CI/CD pipelines, monitoring and observability, Git, Linux, Agile

Frontend & Testing: ReactJS, React Native, Vue.js, Angular; Playwright, automated testing, headless-browser runtime validation

EXPERIENCE

A&C Technology, Inc

Software Development Engineer (AI) · Full-time

Irvine, CA
Dec 2025 – Present

- Design and own the core **LLM generation engine** behind RexTrix (rextrix.ai), an AI-native platform that turns a text prompt into a playable mini-game. Built the **multi-agent orchestration pipeline** on AWS Lambda, Fargate, and SQS (event-driven), with a modular, composable prompt- and context-engineering design supporting 2D, 3D, multiplayer, and motion-controlled modes; serving **100,000+** users and **20,000+** generated games to date.
- Built a **Retrieval Augmented Generation (RAG)** system over user-uploaded and AI-generated asset libraries — embeddings, semantic search, and relevance ranking — that auto-selects the most suitable assets; paired with an original asset-isolation architecture that prevents malformed generations and lets users remix or reskin any game.
- Built a closed-loop **AI evaluation harness** — headless-browser runtime testing with Playwright plus automated design-consistency checks — that benchmarks model output quality and blocks broken or unwinnable games before release.
- Cut per-game generation cost 80%** at equivalent quality via architecture and prompt optimization, caching, and multi-model benchmarking for quality, performance, and cost; instrumented the pipeline with anomaly logging and observability.

A&C Technology, Inc

Software Engineering Intern

Irvine, CA
Jun 2025 – Aug 2025

- Built and maintained the company website on a full-stack architecture: **Java (Spring Boot)** REST backend with a Vue.js frontend.
- Developed **React Native** components shipping new cross-platform features across mobile and web, and added cookie-consent and Google Analytics integrations for the site and web app.

WT Asset Management Ltd

Full-Stack Engineering Intern

Hong Kong (Remote)
Dec 2024 – May 2025

- Led development of a scalable, reliable **ETL data pipeline** leveraging LLMs, a Selenium web scraper, cron jobs, FastAPI, and AWS DynamoDB (NoSQL) to aggregate, analyze, and update structured and unstructured multi-source data daily, delivering timely insights on **800+ companies**; built the transformation and analytics layer with **Pandas and NumPy**.
- Developed an **LLM-powered chatbot with RAG**, using the Zilliz (Milvus) vector database for embeddings and semantic search, integrated with Lark for real-time data-driven responses; ran the pipeline in production for 6+ months.

Ra Labs LLC

Full-Stack Engineering Intern

New York, NY (Remote)
Aug 2024 – Dec 2024

- Implemented a dashboard feature for Legion AI, a **natural-language-to-SQL** BI platform, enabling users to save SQL-executed data visualizations; built both the ReactJS frontend and the NestJS REST API backend.

Uniclone

Project Manager & Software Engineer

May 2024 – Aug 2024

- Led a 3-month team building an AI video-generation SaaS product (Python, Angular, AWS, PostgreSQL) on a **multi-agent framework**.

RockFlow AI

Software Engineering Intern

Aug 2023 – Dec 2023

- Built RockBot, a market-intelligence chatbot using a **LangChain hierarchical multi-agent architecture** with Elasticsearch and RRF hybrid semantic search, improving response speed **40%** and reducing token cost **30%**; optimized TradeGPT with async data pipelines and prompt refinement, lifting news-sentiment prediction accuracy **15%**.

EDUCATION

University of Southern California — M.S., Computer Science, GPA 3.87/4.0 Los Angeles, CA · Dec 2023 – Dec 2025

The Ohio State University — B.S., Computer Science & Engineering, Magna Cum Laude Columbus, OH · Aug 2019 – Dec 2022