Joey (Yepeng) Zhu

me@yepengzhu.com | LinkedIn | GitHub | Portfolio | London, UK

Summary

Backend Engineer specialising in serverless AWS infrastructure, distributed systems, and LLM-powered products — building production-ready APIs and real-time applications from concept to revenue.

Work Experience

Backend Engineer — *Spinnr Tech Ltd, London, UK* Jul 2025 – Apr 2026

- Designed and implemented RESTful APIs for B2B integrations across multiple gaming providers, supporting high-volume transaction workflows with idempotency guarantees.
- Integrated Redis, SQLite and PostgreSQL for reliable data handling and caching.
- Implemented structured logging to improve debugging efficiency and system observability.
- Built a Node.js-based mock platform to simulate external partner APIs, enabling faster integration testing and isolating issues between internal and third-party systems.

Software Development Engineer in Test — *Everbridge, Inc, Beijing, China* Sep 2022 – Mar 2023

- Automated 200+ test cases via Python/Selenium and executed 500+ manual tests per release, reducing testing time by 35% across 31 successful releases.
- Performed API testing using Postman and supported customer technical issues, diagnosing root causes and explaining solutions to non-technical users in English.

Project Experience

Joey OS: Serverless Web Desktop & Interactive Portfolio | Next.js, AWS, Redis, WebSockets, LLM 2026

- Architected an OS-style SPA with Next.js/Tailwind and deployed an event-driven AWS backend (API Gateway, Python Lambda); built a real-time monitoring system powered by DynamoDB and Redis.
- Built a low-latency real-time chat system integrating a web frontend with the Telegram API via WebSockets, achieving sub-second message delivery.
- Built a RAG-style LLM assistant grounded in a DynamoDB knowledge base to prevent hallucination; unknown queries auto-logged for iterative KB updates; API secured with Redis rate limiting.

Zoho Mail AI: Intelligent Email Classification | Python, LLM, OAuth2 2026

- Designed a 3-stage pipeline: rule-based classification → dynamic prompt construction → LLM inference; only ambiguous emails reach the model, cutting API cost by 60–70%.
- Engineered a multi-provider LLM client (DeepSeek/Claude/OpenAI/Groq/Ollama) with a 9-shot decision-tree prompt and OAuth2 Zoho Mail integration for consistent, auditable classification.

AI Code Guide: Streaming Coding Mentor | Next.js, Claude API, TypeScript, DynamoDB 2025

- Decomposed software development into a 4-phase AI pipeline (Requirements, Environment, Code Generation, Automated Testing), each phase driven by a dedicated Claude skill agent streaming structured guidance in real time.
- Built a Next.js streaming LLM gateway with multi-provider support (Anthropic, OpenAI, custom base URL), cookie-based session access control, and per-session request budgeting to prevent token abuse.
- Auto-archives AI-generated files in localStorage for instant download; designed DynamoDB-backed session persistence with auto-compact to summarise older context when approaching token limits.

Skills

Languages:	Python, JavaScript, TypeScript, Go, Lua, C/C++, Java, HTML5, CSS
Frameworks:	React.js, Node.js, Express.js, Next.js, EJS, jQuery, Flask, FastAPI, Playwright, Bootstrap
Cloud & Databases:	AWS (Lambda, API Gateway), Docker, CI/CD, Git, WebSockets, PostgreSQL, DynamoDB, Redis, MongoDB, SQLite
AI & LLM:	Claude API, Codex, Cursor, Claude Code, Prompt Engineering, RAG, LLM Streaming

Certification

AWS Certified Solutions Architect – Associate

Education

University of Nottingham Nottingham, UK
MSc in Advanced Computer Science (Merit Award) Sep 2023 – Dec 2024

Beijing Union University Beijing, China
Bachelor of Engineering in Software Engineering (Top 5%) Sep 2019 – Jun 2023