

YUEZE ZHU

AI APPLICATION ENGINEER / AI FULL-STACK ENGINEER

Beijing, China | +86 172 7777 1041 | 738846162@qq.com
zhuyueze.top

PROFILE

Product engineer with 7+ years of experience delivering web, cross-platform, and SaaS products. I currently own end-to-end engineering for an AI academic-writing platform across the Next.js/TipTap editor, Django/FastAPI services, agent workflows, RAG, AI-service integration, and cloud-native delivery.

I take AI products from requirements and interaction design through service development, tool orchestration, retrieval, deployment, and production diagnostics. My focus is turning LLM output into controlled, verifiable, and recoverable product behavior.

CORE SKILLS

AI Engineering: LLM applications, agent workflows, tool calling, RAG, hybrid search, reranking, prompt design, grounding, SSE streaming, and token governance

Full Stack: Python, FastAPI, Django, Node.js, Fastify, MySQL, PostgreSQL, Redis, Prisma, REST APIs, and WebSocket

Cloud Native: Docker, Kubernetes, Nginx, GitLab CI/CD, health checks, graceful shutdown, CDN delivery, Sentry, and LiteLLM

Frontend: React, Next.js, Vue, TypeScript, Ant Design, TipTap/ProseMirror, Yjs/Hocuspocus, Electron, Flutter, Taro, and uni-app

PROFESSIONAL EXPERIENCE

Beijing Jiansheng Technology | AI Application Engineer / AI Full-Stack Engineer

Jan 2023 - Present

- Own product delivery across web, desktop, browser extensions, and Office/WPS add-ins, including editor, backend, Agent/RAG, AI-service, deployment, and production support.
- Evolved assisted writing from one-shot generation into workflows for outline confirmation, section drafting, rewriting, and full-document operations.
- Designed document-version, node-hash, context-match, atomic-transaction, and snapshot-recovery safeguards for precise AI editing.
- Reduced initial-page resources from approximately 5 MB to 200 KB and improved observability with Sentry, LiteLLM, operation IDs, and analytics.

Beijing Youle Today Technology | SaaS Product Engineer

Nov 2023 - Dec 2024

- Built FastMoss, AdsMoss, and YooFinds for global markets, covering authentication, payment UI, permissions, localization, analytics dashboards, and reusable admin modules.

Remote Project | Mobile and Management Systems

Oct 2022 - Oct 2023

- Delivered a uni-app HR management application for an agricultural business, including feature development, API integration, and release iterations.

Beijing Shanqu Information Technology | Web / Flutter Engineer

Nov 2020 - Sep 2022

- Developed consumer apps, mini-programs, and operations dashboards for KK Keyboard, Shanmeng Emoji, Shanmeng Niuniu, and Rocket Pics.

Beijing Shangrui Technology / Beijing Zhibai Software | Web Engineer

Dec 2018 - Sep 2020

- Built web, mini-program, and admin experiences for ModeSens and Yishanhui, including SSR pages, API integrations, and shared UI components.

SELECTED AI AND FULL-STACK PROJECTS

Qinyan Academic - AI-Assisted Writing and Agent Editing System

qinyanai.com/plugin

NEXT.JS, REACT, TIPTAP, ELECTRON, DJANGO, FASTAPI, REDIS, SSE, RAG, DOCKER, KUBERNETES

- Led frontend architecture across web, desktop, browser, and Office/WPS add-in surfaces; responsibilities expanded into Django/FastAPI services and AI workflows.
- Delivered literature reading, tagging, AI topic ideation, academic Q&A, literature search, knowledge-base retrieval, assisted writing, citation insertion, and multi-platform editing.
- Registered retrieval, citation, and mutation tools dynamically by user intent while isolating concurrent request state.
- Added cancellation, retry, deterministic terminal states, and snapshot recovery for long-running SSE operations.

Precise AI Editing and Document Consistency

TIPTAP/PROSEMIRROR, REACT, TYPESCRIPT, SSE, DJANGO, FASTAPI

- Converted free-form model output into structured selection, paragraph, section, document, and batch mutations.
- Rejected stale output using document versions, node hashes, and context matching; applied approved changes as atomic, undoable transactions.

RAG Knowledge Base and Grounded Citations

HYBRID SEARCH, EMBEDDINGS, RERANKING, REDIS, TOKEN BUDGETS, CITATION GROUNDING

- Consumed upstream rerank results after broad recall, then applied dynamic thresholds, document deduplication, and adjacent-chunk merging.
- Preserved source metadata and enforced evidence-backed citations while balancing retrieval coverage against context cost.

Agent Runtime and AI-Service Integration

PYTHON, FASTAPI, AGENT TOOLS, WORKFLOW SKILLS, SSE, LITELLM

- Built request-scoped tool policies and directly improved model gateways, tool registration, retrieval, citation, streaming events, and completion outcomes.
- Converted model output into frontend-consumable protocols and added regression coverage for request isolation, tool output, and citations.

Shopify Advertising Attribution and Conversion Tracking SaaS

gottracked.io

NEXT.JS, REACT, TYPESCRIPT, FASTIFY, POSTGRESQL, PRISMA, OAUTH, WEBHOOKS, DOCKER

- Independently delivered the dashboard, APIs, database, Shopify extension, containerized deployment, and production handoff documentation.
- Implemented multi-account and multi-store models, OAuth, webhook HMAC, GDPR webhooks, Web Pixel collection, event deduplication, order attribution, retries, and Meta/TikTok/Google payload adapters.

AI Product Engineering and Cloud Delivery

DOCKER, KUBERNETES, NGINX, GITLAB CI/CD, SENTRY, YJS, HOCUSPOCUS

- Improved multi-stage frontend images, Nginx health checks, SSE/WebSocket proxying, and incremental object-storage/CDN publishing.
- Added graceful shutdown behavior that persists active Yjs document state before collaboration-service termination.

ADDITIONAL PRODUCT EXPERIENCE

FastMoss - Global TikTok Analytics SaaS

fastmoss.com

NEXT.JS, REACT, UMI, TARO, ANT DESIGN, I18N, MYSQL, REDIS, NGINX

- Built authentication, third-party payment flows, localization, permissions, analytics pages, and reusable modules for TikTok data across 15+ markets.
- Introduced Next.js SSR for SEO-sensitive pages and reusable tables, filters, charts, and permission components to accelerate product iteration.

ModeSens - Cross-Border Luxury Commerce

modesens.cn

NUXT, KOA2, VUE, BOOTSTRAPVUE, I18N, PYTHON, SCRAPY, MYSQL, REDIS

- Delivered authentication, payment, product detail, account, community, editorial, and sizing experiences for an international luxury-commerce platform.
- Integrated overseas commerce data and extracted reusable modules; later refactored selected features with the Vue Composition API.

Shanmeng and KK Consumer Applications

VUE, NUXT, TYPESCRIPT, FLUTTER, DART, UNI-APP, KOA2, PHP, GO, MYSQL, REDIS

- Delivered GIF search and distribution, creator submissions, memberships, payments, virtual goods, content management, analytics, and mini-program experiences.
- Built reusable mobile payment bridges, visibility-aware countdown modules, media-frame extraction, and layered sticker/drawing composition workflows.

OTHER PROJECTS

AdsMoss: Multi-country advertising analytics SaaS; authentication, payment UI, permissions, localization, dashboards, and reusable data components.

YooFinds: Global TikTok product-sourcing platform; Next.js pages, data presentation, localization, and shared SaaS modules.

Mobile products: Norse HR app, Rocket Pics creator-submission app, and Shanmeng Niuniu digital-collectibles app.

Web products: Yishanhui responsive web and mini-program product; personal portfolio built with Next.js, React, Docker, and Nginx.

EDUCATION

Hebei Vocational College for Correctional Police | Software Technology

Graduated 2017

AVAILABILITY

Available for remote freelance work involving AI applications, agent workflows, RAG, full-stack SaaS delivery, rich-text editors, and production deployment. Comfortable collaborating asynchronously across time zones.