ZHENYU YANG

Toronto, ON · PR · +1 (425) 230-7227 · yangzhenyu528@gmail.com · linkedin.com/in/zhenyu-yang-740b14275 · github.com/JuliusYang3311/verso-agent

PROFESSIONAL SUMMARY

AI Infrastructure Engineer and Full-Stack Developer with **2+ years AWS expertise** building production-grade Generative AI systems. Specialized in architecting autonomous agent platforms, enterprise ML pipelines, and scalable RAG systems, translating cutting-edge AI research into production-ready infrastructure.

AWS CERTIFICATIONS

AWS Certified Machine Learning – Specialty

AWS Certified Cloud Practitioner

PROJECTS

Verso AI — Autonomous Agent Platform | Creator & Lead Architect | Dec 2025 – Present

- **Self-Evolving Engine (GEP):** Designed Genetic Expression Programming pipeline for autonomous code optimization. Extracts runtime signals, generates mutations via AST transformation, validates in sandboxed **Docker** environments, auto-deploys or rolls back. **92% self-healing rate**, **500+ daily invocations**, zero manual intervention.
- **Three-Layer Memory (L0/L1/L2):** Architected progressive retrieval — L0 summaries (~100 tokens) for file pre-filtering, L1 overviews (~500 tokens) for structure, L2 full text on-demand. Hierarchical search with hybrid scoring (vector 0.7 + BM25 0.3), score propagation (α=0.7), adaptive thresholds. **+42% retrieval accuracy** vs. fixed top-k RAG, **3.2× information density**.
- **Async Execution:** Decoupled I/O from agent Turn execution. Turns run as background tasks with steer() injection for new messages. Priority queuing, graceful cancellation, state persistence. **<900ms P95 tool call latency**, **+5× concurrent processing capacity**.
- **Dynamic Context Builder:** Token-budget-aware assembler balancing recent messages vs. retrieved knowledge. Sliding window with exponential decay, automatic summarization. Input tokens controlled within **200k limit**. Conversation rounds **+65%** without session restart.
- **Smart Router:** Real-time complexity scoring routes tasks to optimal models (Flash/Sonnet/Opus). Provider failover with exponential backoff, circuit breaking, cost-aware load balancing. **40% token cost reduction**, **<300ms P95 latency**.

OpenClaw Contributor (PR #9123) | Open Source

- Designed smart routing system with complexity-based model selection, provider abstraction supporting 10+ LLM providers, unified error handling. Contributed plugin architecture. Adopted by 15+ contributors, serving 1M+ monthly requests.

Independent AI Engineer / Consultant | Freelance | Jan 2025 – Present

- **RAG Search Service:** Built production hybrid retrieval (**OpenSearch** vector + BM25) with **Redis** embedding cache, async batch indexing, cross-encoder re-ranking. Optimized index sharding for 100K+ docs. **<300ms P95 latency**, **89% precision**. Reduced infrastructure costs 35%.
- **Serverless Pipelines:** Designed event-driven architectures (**AWS Lambda + Step Functions + EventBridge**) with DLQs, exponential backoff, circuit breakers, idempotency keys. Auto-scales **0 to 10K+ concurrent invocations**. 99.9% uptime, zero-maintenance.
- **LLM Integration:** Deployed **AWS Bedrock** models with custom prompt engineering, A/B testing framework, structured output validation. Improved client engagement **+25%**, reduced hallucination rate 40%.

EXPERIENCE

Ji Heng Xiang | Founder & Full-Stack Engineer | Jan 2025 – Present | jihengxiang.com

- Founded and built full-stack web application with **AWS** cloud infrastructure, **CI/CD** pipeline, user authentication, and session management. Manage platform operations and business development.
- Developed LLM-powered chatbot for mysticism consultation with custom **RAG** knowledge base for divination and spiritual guidance. Provide end-to-end client consultations leveraging AI-enhanced conversational interface.

INRIA (French National Institute for Research in Digital Science) | Research Intern | Jul – Dec 2023

- Built high-fidelity C++ agent-based simulation (**R^2 = 0.93** vs. wet-lab data). Implemented spatial hashing (O(1) lookup), cache-aligned memory, SIMD-friendly structures. Profiled with perf, optimized hot paths — **5× runtime reduction**. Scaled to 1M+ agents with real-time visualization.

TECHNICAL SKILLS

Domain	Technologies
Cloud (AWS)	Lambda, Step Functions, Bedrock, SageMaker, Glue, Athena, EventBridge, IAM, SAM
LLM & Agents	Agentic Workflows, Function Calling, Tool-Use, MCP, ReAct/CoT, RAG, Self-Evolving Systems (GEP), Prompt Engineering
Infrastructure	Vector DBs (sqlite-vec, OpenSearch, Pinecone), Hybrid Retrieval, Dynamic Context, Model Routing, Async Execution
Languages	TypeScript, Python, C++, SQL
Tools	FastAPI, Docker, Git, CI/CD (GitHub Actions), PyTorch, Scikit-Learn

EDUCATION

Ghent University | M.Sc. Bioinformatics, **Distinction degree** | 2022 – 2024 | Ghent, Belgium

Ocean University of China | B.Sc. Mathematics | 2018 – 2022 | Qingdao, China