

ZHENYU YANG

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

SUMMARY

Backend & Systems Engineer with 3+ years designing high-throughput distributed systems and performance-critical applications. Creator of Verso — an autonomous agent runtime with async execution decoupling (**<900ms P95 tool calls, +5x concurrency**), three-layer hierarchical retrieval (**+42% accuracy, 3.2x density**), and self-evolving optimization (**92% self-healing rate**). Deep expertise in serverless architecture (AWS Lambda, Step Functions, EventBridge), high-performance C++ optimization (spatial hashing, cache alignment, SIMD), and scalable search infrastructure (hybrid retrieval, embedding caching, query optimization). Proven track record building production systems serving 1M+ monthly requests with 99.9% uptime. AWS Certified ML Specialty.

EXPERIENCE

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

- High-Performance Search:** Architected 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%.
- Event-Driven Serverless:** Designed production pipelines (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.
- API & Integration:** Built RESTful APIs (FastAPI) with request validation, rate limiting (token bucket), response caching (Redis). Integrated AWS Bedrock with streaming, timeout handling, graceful degradation. Improved client engagement **+25%**.

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

- Simulation Engine (C++):** Designed high-fidelity agent-based simulation (**R² = 0.93** vs. experimental data). Implemented spatial hashing (O(1) lookup), cache-aligned memory, optimized hot-path allocations. Profiled with perf/valgrind, applied SIMD vectorization. **5x runtime reduction**, scaled to 1M+ agents with real-time visualization.

SYSTEMS & INFRASTRUCTURE PROJECTS

Verso — Autonomous Agent Runtime | Creator & Lead Architect | Jan 2025 – Present

- Async Execution Architecture:** Decoupled message 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, +5x concurrent processing capacity**.
- Three-Layer Hierarchical Retrieval (L0/L1/L2):** Progressive loading — L0 summaries (~100 tokens) for file pre-filtering, L1 overviews (~500 tokens) for structure, L2 full text on-demand. Hierarchical search: Phase 1 file-level hybrid scoring (vector 0.7 + BM25 0.3) → Phase 2 chunk-level with score propagation ($\alpha=0.7$), adaptive thresholds. **+42% retrieval accuracy** vs. fixed top-k RAG, **3.2x information density**.
- Self-Evolving Optimization (GEP):** 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.
- 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.

OpenClaw Core Contributor (PR #9123) | Open Source

- Architected smart routing system with complexity-based model selection, provider abstraction supporting 10+ LLM providers, unified error handling with retry/failover. Designed plugin interface for community extensions. Adopted by 15+ contributors, serving 1M+ monthly requests with 99.95% success rate.

Multi-Agent Orchestration System | Sep – Dec 2024

- Designed stateful 3-agent pipeline with coordinated handoffs, shared state management (Redis), human-in-the-loop checkpoints for automated business analysis. Message passing with retry logic, timeout handling, partial result recovery. Reduced analysis time from 8 hours to 45 minutes, 91% accuracy vs. expert analysis.

TECHNICAL SKILLS

Domain	Technologies
Languages	TypeScript, Python, C++, SQL
Backend & APIs	FastAPI, Event-Driven Design, Async Execution, REST APIs, WebSockets, Process Supervision
Cloud (AWS)	Lambda, Step Functions, EventBridge, Bedrock, SageMaker, Glue, Athena, IAM, SAM, CloudWatch
Infrastructure	Docker, CI/CD (GitHub Actions), sqlite-vec, OpenSearch, Pinecone, Hierarchical Retrieval, Sandbox Isolation
Performance	Spatial Hashing, Cache Optimization, Memory Access Patterns, Async Pipelines, Concurrency Control, FD Management
AI/ML	LLM Integration, RAG Pipelines, Multi-Agent Systems, PyTorch, Scikit-Learn

EDUCATION & CERTIFICATIONS

Ghent University | M.Sc. Bioinformatics (Distinction) | 2022 – 2024 | Ghent, Belgium

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

AWS Certified Machine Learning – Specialty | AWS Certified Cloud Practitioner