Zaid Yusuf

ML Platform & AI Infrastructure Engineer | LLM Inference & Serving | CKA

zaidyusuf.com | github.com/Yusufteppei | linkedin.com/in/zaid-yusuf-de |
yusufteppei11@gmail.com | Open to remote or local roles

Summary

ML platform and infrastructure engineer with ~4 years of experience building cloud-native systems for machine learning, backend services and deployment pipelines. Certified Kubernetes Administrator with in-depth study of LLM inference and serving (KV caching, batching, parallelism, GPU capacity planning, SLO-driven observability) and hands-on research implementation experience in PyTorch and JAX. Strong systems fundamentals: built a relational database engine from scratch in C. 2x IMO contestant.

Professional Experience

Machine Learning Engineer, Digprom (Contract, part-time) | Jun 2023 - Present

- Built and maintained machine-learning workflows and supporting data pipelines for government and university client projects at an IT consulting firm.
- Developed backend systems supporting data processing and ML applications.
- Applied software engineering practices to improve reliability, reproducibility and maintainability.
- Worked with engineering and domain teams to translate requirements into technical systems.

Full-Stack Engineer, Turing (Contract) | Jun 2024 - May 2025

- Built and maintained backend systems and APIs supporting AI training projects.
- Worked with cloud infrastructure and production deployment workflows, mostly Docker-based.
- Designed reliable application architectures and data-driven services.

Platform & Machine Learning Engineer, 9ijakids / Lorem Excellentiam (Contract) | Dec 2023 - Jul 2024

- Built cloud-native infrastructure supporting scalable machine-learning systems [add: cloud provider, orchestration tools].
- Developed production ML solutions and inference workflows.
- Automated deployment and experimentation pipelines.
- Improved reliability, reproducibility and monitoring: adding monitoring cut average incident response time to one-fifth of its previous level and made it clear which team owned each incident.

Founding Engineer, Edulethics Africa (Contract) | Jan 2023 - Dec 2023

- Led engineering strategy, software architecture and technical decisions across product development.
- Designed and implemented the platform's core architecture, serving over 45,000 users.

Selected Projects

LLM Inference & Serving Study | zaidyusuf.com/projects/ml-systems

- Implementation-led study of the systems choices behind reliable LLM applications: prefill/decode architecture, KV-cache management, continuous and paged batching, prefix caching, speculative decoding, quantization, tensor and pipeline parallelism, GPU/VRAM capacity planning and model routing.
- Covers serving metrics and observability: TTFT, TPOT, tokens/sec, p50/p95/p99 latency, KV-cache utilization, health checks, failure modes and SLO-driven scaling.
- [add: measured results, e.g. throughput and latency numbers from a vLLM benchmark you run]

FensterSQL: relational database engine | C, Flex/Bison | github.com/Yusufteppei/fenstersql

- Built from scratch: SQL parsing, DDL/DML support, page-based storage, B+ tree indexing, catalogs and persistent metadata, disk-backed binary I/O, SQL shell, with manual memory management.

BioAgent: tool-calling agent | LLM fine-tuning, SFT/LoRA | github.com/Yusufteppei/bio-agent

- Fine-tuned open-source LLMs with SFT/LoRA to build a biology-focused tool-calling agent. Live demo: zaidyusuf.com/projects/bio-agent

- Studying agent reliability: tool boundaries, idempotency, retries and failure handling.

Research implementations in PyTorch | github.com/Yusufteppei

- ProtDiffusion (diffusion over SE(3)/SO(3) protein backbone geometry), AlphaFold2-mini (Evoformer, triangle attention, Invariant Point Attention) and scGPT-mini (single-cell foundation model), each implemented from first principles.

Technical Skills

- **Infrastructure:** Kubernetes (CKA), Docker, AWS, Linux, CI/CD, PostgreSQL, cloud infrastructure
- **ML systems (current focus):** LLM serving, KV caches, continuous/paged batching, prefix caching, speculative decoding, quantization, distributed inference, GPU/VRAM planning, observability and SLOs
- **ML and languages:** PyTorch, JAX, Hugging Face, Python, C, C++, TypeScript, SQL, Bash
- **Backend and web:** FastAPI, Django, Next.js, React

Education & Honors

- Nile University of Nigeria, Bachelor's coursework in Computer Science | Sep 2017 - Jan 2020
- International Mathematical Olympiad (IMO): 2x contestant (2015, 2016)
- Pan-African Mathematics Olympiad (PAMO) 2015: Bronze medal
- Cowbell Mathematics Competition 2014: 2nd place nationally (silver)
- Certified Kubernetes Administrator (CKA)
- Languages: English; German (B2)