Xushuo

北京大学 PEKING UNIVERSITY

LLM Application Backend / Backend Engineering

xushuo_1471@foxmail.com

GitHub | Homepage

Education

Peking University Sep 2021 - Jul 2026

B.S. in Computer Science and Technology School of Electronics Engineering and Computer Science

Projects

Local Multi-Tenant RAG Backend | FastAPI, PostgreSQL/pgvector, Redis, MinIO | GitHub

- Built for enterprise knowledge-base Q&A: enforced tenant, knowledge-base, and document permissions during full-text and vector retrieval so unauthorized content never enters results; added versioned idempotent indexing and lease-based worker recovery
- On the public SciFact dataset (5,183 documents and 188 human-labeled queries), the top five results covered about 86.5% of relevant evidence (Recall@5 = 0.865); this is an offline evaluation, not production-traffic evidence

LLM Evaluation & Prompt Regression Platform | FastAPI, SQLAlchemy, Vue 3, GitHub Actions | GitHub

- Built to compare models, prompts, and evaluation settings before releases introduce quality, cost, or safety regressions; versioned datasets, targets, and evaluators with traceable records, plus an independent worker with retries, caching, and interruption recovery
- Compared four configurations on a fixed 100-case deterministic local regression set, producing 400 traceable evaluation records and replaying 100/100 cached cases; quality, cost, latency, and safety can be gated independently

Secure Business Workflow Agent | FastAPI, PostgreSQL, MCP | GitHub

- Built for workflows where models assist with approvals and writes without executing risky actions directly: the model only proposes operations, while the server enforces schema, RBAC, and tenant checks and routes high-risk writes to durable human approval
- Handled timeouts, cancellations, and duplicate execution with PostgreSQL checkpoints, an outbox, and idempotency keys; 138 automated tests cover authorization bypass, injection, approvals, and recovery at 92.65% branch coverage; the public demo is clearly labeled as static

TinyLlama CUDA Inference Engine | C++20, CUDA, CMake, cuBLAS | GitHub

- Built to study and optimize small-model inference on a consumer GPU: implemented TinyLlama-1.1B CPU reference, CUDA FP16/W8A16 execution, KV cache, and Batch-4 scheduling; completed 249 kernel checks on an RTX 3080 Laptop GPU and passed memory and race checks
- At a fixed context length of 128 with 32 generated tokens, Batch 4 reached 61.35 tok/s aggregate throughput, about 3x Batch 1 (20.48 tok/s); W8A16 cut peak GPU memory by 40.4% but reduced throughput by 12.5%, demonstrating the memory-performance tradeoff

Skills

- **Languages**: Python, C++, CUDA, SQL
- **Application Development**: FastAPI, PostgreSQL/pgvector, Redis, MinIO, Docker, Vue 3
- **Engineering**: Linux, Git, GitHub Actions, OpenTelemetry, Prometheus
- **Computer Science**: Data Structures & Algorithms, Operating Systems, Compilers, Computer Architecture, Databases
- **English**: CET-4 582; proficient in reading technical documentation and research papers