AARON CHEN

(216) 801-3618 | aaronbychen@gmail.com | linkedin.com/in/aaronbychen | github.com/aaronbychen | aaronbychen.com

EDUCATION

Washington University in St. Louis **Aug 2023 - May 2027**
B.S., Computer Science and Mathematics (GPA: 3.95/4.00) *St. Louis, MO*
- **Achievements:** AIME Qualifier (5% International), Llama Stack Finalist (Meta/8VC), ARML National Bronze, ACSL State 1st

EXPERIENCE

Portal **May 2025 - Aug 2025**
AI Software Engineering Intern *Houston, TX*
- Built LLM-powered career assessment & recommendations using **LangChain + LLM APIs** on **Node/Express** with **Postgres/Prisma** persistence; improved user satisfaction **+80%** across **5** tracked metrics (activation, retention, completion, NPS-style feedback)
- Implemented resume ingestion (PDF/DOC/DOCX to **deterministic JSON**) via **temperature=0** parsing, schema validation, and **retries + exponential backoff**; added **token/cost telemetry** + budget alerts to keep spend predictable and observable under load
- Developed a hybrid ATS scoring engine combining **rules + TF-IDF/cosine (natural)** with **LLM rewrite guidance**; produced **explainable score breakdowns** (format/content/keywords) and auto-generated targeted edits, improving recommendation quality
- Hardened LLM-backed APIs with strict **timeouts (Promise.race)**, **bounded retries**, and circuit-style fallbacks; integrated a DB-backed **credit ledger** (UserCredits/CreditTransaction) to gate usage, audit spend, and prevent long-hanging requests from cascading
- Shipped secure multi-tenant onboarding: **SAML SSO (passport-saml)** with dynamic per-domain configs + metadata, plus **JWT access/refresh** in **HttpOnly** cookies; enforced **RBAC** via InstitutionMembership roles for institution-scale authorization
- Automated content ingestion with **Python + Selenium** and **Google Scripts**: compiled **33k+** career videos, processing **1k+/week** across **200+** professions; expanded library **+400%** and reduced manual ops via scheduled runs, deduping, and failure handling
- Operationalized AI coding workflow with **Claude Code**: ran multiple agent processes in parallel using **rules/subagents/hooks** and **least-privilege permissions**; accelerated development **~2-3×** and pioneered automated pre-prod code review to catch bugs earlier

CN Business Herald **May 2024 - Aug 2024**
AI Software Engineering Intern *Beijing, CN*
- Owned the end-to-end **RAG** platform: document ingestion/cleaning, chunking, **Azure Cognitive Search** indexing & retrieval, and **LangChain** orchestration (prompt templates, citation formatting, guardrails) for offline news analysis
- Built a self-serve **on-prem** admin dashboard (SSO, streaming, **MongoDB**-backed persistence) for retrieval tuning and output evaluation; cut combined platform run costs + ops/support time by **~27%** (normalized per 1k requests, rolling averages)
- Integrated **RouteLLM** routing to local models behind an **OpenAI-compatible API gateway**; reduced cloud API-served inference by **~38%** (gateway logs + token ratios + billing) while keeping client integrations unchanged
- Productized two editorial workflows (policy impact briefs; company/topic dossiers with timelines + Q&A banks), shrinking research from **0.5-2 days** to ~**10-20 mins** and standardizing citation-backed outputs for **200+ users**

PROJECTS

E-Commerce Recommendation Engine | ***PyTorch, Faiss, TorchServe, Kubernetes, Spark, Redis, XGBoost***
- Built a two-stage recommender (recall → ranking) using **PyTorch + Faiss** on **10M+ interactions**; evaluated with **Recall/NDCG** and tuned locally for data sensitivity and reproducible offline metrics
- Served a **Two-Tower DNN** via **TorchServe** on **Kubernetes** with batching/autoscaling, achieving **p99 9ms @ 1.2k RPS/GPU** on clickstream events (views/add-to-cart/purchases) under steady load
- Streamed per-session features with **Spark Structured Streaming** (1s micro-batches) and cached in **Redis** for low-latency online scoring; ranked top-K with **XGBoost/LambdaMART** and delivered end-to-end real-time ranking

NL Video Scanner (Llama Stack Finalist) | ***Llama-Vision, LlamaIndex, NoSQL, OpenCV, FastAPI, Docker***
- Delivered a privacy-preserving **natural-language** video scanner at **1 FPS** using **Llama-Vision** + **OpenCV**; achieved **85%+ Hit@5** on 100-300 clips with **zero data egress**
- Built local retrieval/indexing (**LlamaIndex + NoSQL**) for frame embeddings + metadata; optimized fast **top-k** queries
- Enhanced robustness with OpenCV preprocessing (frame sampling, denoise, **CLAHE**, color normalization) for low-light/high-contrast scenes, improving retrieval quality by **40%+** and reducing false positives via tunable thresholds
- Served inference via **FastAPI** (async streaming + batching); improved concurrency and cut end-to-end latency by **~50%** (load-tested)

SKILLS

- **Languages:** Python, TypeScript, JavaScript, SQL, Java, C/C++
- **AI/ML:** PyTorch, TensorFlow, scikit-learn, Transformers, RAG, Faiss, LangChain, LlamaIndex, OpenCV, NumPy, Pandas
- **Technologies:** AWS, Azure, Node.js, Express.js, FastAPI, Spring Boot, React, PostgreSQL, MySQL, Prisma, MongoDB, NoSQL, Redis, Kafka, Elasticsearch, Docker, Kubernetes, GitHub Actions, Socket.IO, Etcd, Spark, Hadoop, Linux, Git