Portfolio - Backend & AI Agent

Yizhou Shen / shen.yizhou@foxmail.com / Nanjing

Trading API performance (internship, 2026.07-2026.10)

- Two-level cache on order detail: hit rate 43% to 91%, origin QPS 1800 to 160.
- p95 from 380 ms to 95 ms; 2.1M requests served per day.
- 17 full-table scans fixed with EXPLAIN and 4 composite indexes; slow queries from 23K to 900 per day.
- Statement export moved to async task: timeout error rate 8.4% to 0.2%.

Agent tool layer (internship, 2026.01-2026.05)

- 9 function tools with JSON Schema validation and one retry: success 79% to 94%.
- Built a 160-case eval set; 21 regressions caught before release in three months.
- Heading-aware chunking plus BM25/vector fusion: Recall@5 from 61% to 84%.

Agent eval checklist

1. Failure type is labelled (wrong tool / hallucinated args / loop / gave up)
2. Eval cases cover long-tail tools, not only the ones I changed
3. Ground truth reviewed by a second person
4. Tool-call success counted per call, not per conversation
5. Retry limit and timeout set explicitly
6. Fallback path works when the tool keeps failing
7. Latency p95 recorded together with success rate
8. Regression run before every release