MLI RECRUITER SIGNAL

ML Infrastructure

Retrieval, inference serving, and distributed training.

AARYAN PATEL

Systems and ML Infrastructure Engineer

aaryansp26@gmail.com

asp53826.github.io

github.com/asp53826

THREE SYSTEMS TO INSPECT

01 annlite

C++17 + Python

1.83x

FAISS QPS AT 0.999 RECALL

51 tests; at 98.5% recall it performs 26x fewer distance computations than exact scan.

FAILURE BOUNDARY

FAISS remains 10-20% faster in the low-recall regime; the crossover appears only at high recall.

START `git clone https://github.com/asp53826/annlite && cd annlite && pip install -r requirements-...`

02 vllm-lite

Python

94%

KV SLOT UTILIZATION VS 21%

90 tests; paged scheduling reaches 2.9x sequential throughput and 5.5x better median TTFT.

FAILURE BOUNDARY

CPU-scale simulator with a small model; it does not claim production GPU kernel performance.

START `git clone https://github.com/asp53826/vllm-lite && cd vllm-lite && pip install -r requireme...`

03 dist-train

Python

58.7 MB

PER WORKER AT EIGHT WORKERS

57 tests; the naive baseline sends 234.9 MB per worker under the same eight-worker case.

FAILURE BOUNDARY

Loopback gloo on a 10-core machine; efficiency falls past four workers and small tensors favor naive collectives.

START `git clone https://github.com/asp53826/dist-train && cd dist-train && pip install -r require...`

VERIFICATION CONTRACT

- Claims name the comparison baseline and units.
- Commands reproduce the public correctness path.
- Failure boundaries remain beside the measured result.
- The evidence manifest connects claims to methods and sources.