

Wenjie Luo

AI Inference and Model Evaluation Engineer | Graduating Jan 2027

Singapore | +65 8084 8140 | e1576499@u.nus.edu
github.com/LUOaini1213 | luoini1213.github.io

PROFILE

Evaluate inference performance and model behavior through controlled experiments and numerical checks. Hands-on experience with SGLang/vLLM serving, GPU kernels, PyTorch training and scene-level analysis.

EDUCATION

- National University of Singapore | MSc Civil Engineering**Aug 2025 - Jan 2027
- Transport focus; research in transportation information and control.
- Chang'an University | BEng Transportation Engineering**Sep 2021 - Jun 2025
- GPA 3.16/5.0; approximately top 15% of the cohort.

PROJECTS

- SGLang/vLLM Serving and Performance Analysis**Sep 2026
- On a T4 with Qwen2.5-0.5B, traced SGLang's throughput drop to decode-graph capture stopping at batch 8; tested causality by disabling graphs and extending capture.
 - Raised the capture limit to 32: batch-32 throughput rose from 608 to 2,690 tok/s (4.42×) in the same experiment. Logged parameters and one measurement per batch.
- fp16x3 Transformer GPU Optimization**Aug 23, 2026 - Present
- Wrote Triton kernels to split operands into fp16 high/low components for one expanded-K cuBLAS Tensor Core GEMM; used fused attention and chunking to reduce memory use.
 - T4: 13/13 FP32 shapes passed (max abs error 9.5e-6); 3 independent runs/shape gave 2.83× median speedup over the FP32 reference. A separate FP16 100k run peaked at 14.2 GB.
- NAVSIM | Planning Model Training and Evaluation**Aug - Sep 2026
- Trained MLP planners on 103,288 navtrain scenes and evaluated 12,146 navtest scenes; used a 2×2 map/agent-input design, three seeds per cell and paired 95% confidence intervals.
 - Reproduced PDMS baselines: ConstantVelocity 20.7 (paper 20.6), Human 94.6 (94.8). Adding ground-truth boxes significantly lowered map-aware models' PDMS across three seeds despite similar L1.

EXPERIENCE

- Far East Facade | Project Engineer, Engineering Department**Jun 2026 - Present
- Used Python to extract delivery batches and test results from supplier emails, supporting status tracking for 15 façade material types.
- Caixun Technology | Product Intern**Jun 2025 - Aug 2025
- Moved the hard-to-find prize-draw entry from the page bottom to a side pop-up to improve visibility; defined requirements, designed interaction prototypes and followed development and testing.
 - Independently wrote 3 PRDs and prototyped 5 core pages in Axure; proposals were accepted and scheduled for development. Tracked fixes and retesting, helping resolve over 10 issues.

SKILLS

Inference: SGLang, vLLM, CUDA C++, Triton, PyTorch, CUDA Graphs
Evaluation: controlled experiments, numerical checks, paired bootstrap, factorial designs, multiple seeds