

SOOBIN PARK (EVAN PARK)

LLM Inference & Serving Engineer

subin1031@gmail.com | +82-10-9199-9786 | Seoul, Korea | cv.subin.rs | github.com/subin9 | linkedin.com/in/soobinpark1

SUMMARY

LLM serving engineer fixing throughput, memory and accuracy problems at the runtime and kernel level on NVIDIA GPUs, AMD MI300X and Rebellions NPUs. **Forward Deployed Engineer at Rebellions; committer on vLLM Semantic Router** and lead of its Router Models & Inference Runtime working group.

EXPERIENCE

Forward Deployed Engineer | Rebellions (AI accelerators) | Seongnam June 2026 – Present

- **Memory-budget helper.** Auto KV sizing left out the sampler, which compiles later in another process, so 70 graphs fell back to CPU on Qwen3-8B with no error; greedy benchmarks hid a 4× throughput loss. Built a helper that reserves sampler memory before the block search (0 fallbacks). FDEs use it for deployment specs.
- **Dynamic batching slowdown.** Dynamic batching filled memory by binary search, so prefix-cache hits failed the KV copy and whole buckets fell back to CPU. Re-measured with my fix: **2.4–2.7×** throughput at the same batch size.
- **Prefill/decode disaggregation.** NPU prefill ran one request per step and stalled decode (ITL spikes of 78–86×). On 16 NPUs, disaggregation held ITL within 2.7× of baseline, but no static P:D split beat aggregated serving on one node. Enabled CMA in NIXL/UCX for 7.4× faster KV transfer.
- **Model porting.** PaddleOCR-VL on the NPU within 0.2 points of the published OmniDocBench v1.6 score; 12.7× throughput after fixing CPU thread contention.

OPEN SOURCE

vLLM Semantic Router | Committer, Lead of Router Models & Inference Runtime WG Aug 2026 – Present

The routing layer of the vLLM ecosystem. 98 PRs, 83 merged; 9 PRs reviewed.

- **Decision 3.0 (27B) runtime optimization and quantization.** Captured the noncausal forward as CUDA graphs: 4× lower latency on d3-lite ([#4999](#)). Built and published FP8 and AWQ int4 packages ([d3-FP8](#), [d3-W4A16-AWQ](#)) and made the runtime load them as stored, with my Triton FP8 activation kernel ([#5066](#)). FP8 uses 1.6× less memory and keeps **99.08%** of BF16 answers. On MI300X it is slower than BF16 without graphs and 1.26× faster with them ([#5270](#)).
- **CUDA graphs for Vela 2.0.** Switching each fast path off on MI300X showed Vela 2.0 never replayed a graph ([#4811](#)). Per-shape capture cut latency 3.8× on the 0.8B (4B: 1.9×), bit-exact ([#4813](#)).
- **Runtime robustness.** Per-tensor checkpoint staging cut host memory 8× while loading the 27B model ([#5184](#)). An out-of-memory forward now answers 413, or a retryable 429, instead of 500 ([#5183](#)).
- **Long-context classifiers.** Guardrails truncated input at 512 tokens with no warning. Fixed six call sites with overlapping windows and moved attention to a chunked kernel: ~16× less memory at 8K tokens ([#3206](#), [#3402](#)).
- **Evaluation.** A label-order mismatch cost **27 points** on MMLU-Pro ([#3247](#)); 57% of a benchmark was training data ([#4308](#)).
- **Hugging Face transformers** — co-authored an FP8 weight-scale loading fix ([#47152](#)); when it broke mislabeled checkpoints, wrote the revert ([#48072](#)) and reported 42 mislabeled Hub repos, which Red Hat relabelled.
- **Ollama** — whole-prompt batch for noncausal decision-model GGUFs, which failed past 512 tokens ([#19159](#), open).

EARLIER EXPERIENCE

Solution Engineer Intern | Microsoft Korea | Seoul Jan 2026 – Mar 2026

- Technical lead for a Copilot helpdesk agent and admin dashboard; supported enterprise Copilot PoCs.

Data Scientist Intern | Munich Re Korea | Seoul Jan 2025 – Dec 2025

- Built Treaty Hub alone, a closed-network GenAI platform (RAG Q&A, VLM parsing); presented to 4+ APAC offices.

AI Researcher Intern | NCSOFT Audio AI Lab | Pangyo Mar 2024 – Aug 2024

- PyTorch voice-conversion and SFX-generation prototypes; a sound-tagging tool for sound designers.

EDUCATION & AWARDS

Korea University | B.S. Computer Science, B.A. Korean Language & Literature 2018 – 2025

- GPA 4.17 / 4.50 (CS 4.29). KISTI Director's Award (2024), NIA Director's Award (2023) in national AI competitions.

SKILLS

- **Serving:** vLLM, NIXL, continuous batching, prefix caching, P/D disaggregation, CUDA graphs, FP8 / int4 quantization
- **Hardware & languages:** NVIDIA GPUs (CUDA), AMD MI300X (ROCm), Rebellions NPUs; Python, Rust, PyTorch