Immanuel Peter

ipeter@uchicago.edu | (479) 257-3842 | linkedin.com/in/immanuel-peter | github.com/immanuel-peter | ipeter.dev

EDUCATION

The University of Chicago – **BS Computer Science, BA Physics** | *expected June 2028*

EXPERIENCE

Member of Technical Staff Intern | Tensormesh | *March 2026 – Present*

- Integrated Arize Phoenix and emitted OpenInference traces for all router requests, with SDK and CLI tools for trace/span inspection.
- Built an orchestrator that asynchronously pins prompts in LMCache during inference, exposed through the API, SDK, and CLI.
- Added upstream LMCache support for listing pinned entries and tracking per-key access in the coordinator.
- Led an internal research effort on key weighting for long-context inference.

Software Engineer Intern | Quantum Rings | *June 2025 – August 2025*

- Reduced API request latency by designing queue-driven telemetry processing that decoupled heavy operations from the API, enabling horizontal scaling (AWS SQS, TypeORM)
- Built full-stack admin analytics dashboards tracking user growth and execution volume with SQL time-bucket aggregation and time zone safe filtering (Next.js, NestJS, Recharts)

OPEN SOURCE

LMCache - Python | github.com/LMCache/LMCache

- Added a coordinator endpoint to list pinned cache entries. #4960
- Added per-key access tracking to the coordinator key directory. #4873

Brev CLI – Go | github.com/brevdev/brev-cli

- Authored rsync-first transfers with SCP fallback. #423

vLLM Production Stack - Kubernetes | http://github.com/vllm-project/production-stack

- Fixed default runtimeClass to nvidia in VllmRuntime operator; regenerated CRD manifests. #974
- Added extraVolumes and extraVolumeMounts to routerSpec in Helm chart for custom mounts. #975

PROJECTS

Vision Tower Bench | ipeter.dev/blog/vision-tower-bench

- Froze and benchmarked 16 ViT encoders from multimodal LLMs (Gemma 4, Inkling, DeepSeek V4.1, Nemotron 3.5, Kimi K3, and more) against DINOv3 and SigLIP2 baselines. Ran probes on classification, monocular depth, surface normals, dense correspondence, and occlusion robustness.
- Found geometry performance peaked before final-layer in 23/24 task–tower–capacity combinations. Features after multimodal projector beat raw encoder outputs on 16 of 16 DIODE comparisons.
- Extracted and published 13 standalone vision towers on Hugging Face, verified against parent checkpoints to BF16 tolerance.
- Built eval pipeline: model adapters, sharded feature cache, linear and attention-pooling probes, randomized-SVD PCA for capacity-matched heads, per-cell LR search over 3 seeds. Added label-efficiency and KITTI driving-depth transfer studies.

AutoMoE — MoE Self-Driving Model | github.com/immanuel-peter/self-driving-model

- Trained a 53M parameter modular self-driving stack in PyTorch with 4 expert networks, top-2 gating, and 10-waypoint trajectory prediction.
- Released ~145K CARLA frames across multimodal and multi-camera datasets, and trained/fine-tuned the pipeline with PyTorch DDP on up to 8 GPUs.

CARLA Autopilot Datasets | hf.co/immanuelpeter/datasets | 20K+ Downloads

- Released two MIT-licensed CARLA autopilot datasets (~145K frames, 567 GB) on Hugging Face with 2.2K and 6.5K downloads each; flagship split includes synchronized 4-camera RGB, 32-channel LiDAR, semantic segmentation, and 2D bounding boxes across 24 weather/map scenarios
- Designed run-based train/val/test splits to eliminate temporal leakage, supporting perception, sensor fusion, imitation learning, and RL benchmarking under extreme conditions

Hostess — Docker Compose for Production | hostess.sh

- Engineered a Go-based multi-cloud PaaS control plane that compiles a declarative config into production Kubernetes workloads