Seyfal Sultanov

seyfal.s@outlook.com • (312) 546-4655 • seyfal.com • github.com/seyfal • linkedin.com/in/seyfal • Scholar

Experience

Eagle AI | *Principal Member of the Technical Staff* Aug 2026 – Present

- **EAGLE-4:** Led the open-source release (1.2K+ stars); up to 1.62× faster than MTP, 1.95× than DSpark.
- **EAGLE-Harness:** Created a runtime that makes MTP, DSpark and DFlash 2 up to 1.32× faster and never slower than plain decoding at any batch size.
- **Training infrastructure:** Raised draft-head training MFU 7× over public TorchSpec; run the B300 GPU cluster as code (Slurm, Ansible).
- **Research:** Drive work on new speculative decoding architectures beyond EAGLE-4.

Oxia AI | *Head of Engineering (from Apr 2026), Research Engineer* Dec 2024 – Jul 2026

- **Two-party (MPC) inference:** Led research on a GPU engine that runs Llama-3.1-8B with under 1.5 s time-to-first-token, up to 626× faster than prior CPU-based MPC systems.
- **Encrypted inference gateway:** Built it in Rust; handles 10K+ concurrent AES-256-GCM requests.
- **Equivariant Encryption:** Co-authored the paper; built its production pipeline.

UIC, Electronic Visualization Laboratory | *Part-Time Research Assistant* Aug 2024 – Nov 2025

- **MAGUS (SC25):** Co-developed a runtime that cuts CPU energy by up to 27% with <5% GPU slowdown.

Ashling Partners | *Applied AI Research Intern* Jun 2024 – Aug 2024

- **Computer vision:** Shipped an active-learning pipeline for engineering diagrams (87% accuracy).
- **LLM agents:** Built an agent framework for documents; trained production demand-forecasting models.

Selected Publications

Report '26 **EAGLE-4: One Flexible Drafter for All Batch Sizes and Hardware**
EAGLE-4 Team (incl. **S. Sultanov**). Eagle AI technical report.

npj '26 **Unsupervised Defect Clustering in Atomic-Resolution Microscopy using a CVAE**
R. A. W. Ayyubi, **S. Sultanov**, J. P. Buban, R. F. Klie. *npj Computational Materials*.

SC '25 **Minimizing Power Waste in Heterogenous Computing via Adaptive Uncore Scaling**
Z. Zheng, **S. Sultanov**, M. Papka, Z. Lan. *SC25* (21.2% acceptance rate).

Small '25 **Robust Spectral Anomaly Detection in EELS Spectral Images via 3D Convolutional VAEs**
S. Sultanov, R. A. W. Ayyubi, J. P. Buban, R. F. Klie. *Small*.

arXiv '25 **Encrypted Large Model Inference: The Equivariant Encryption Paradigm**
J. P. Buban, . . . , **S. Sultanov**, et al. arXiv:2502.01013.

Education

B.Sc. (Honors), Computer Science | *University of Illinois Chicago* Jan 2022 – Dec 2024
GPA 3.9/4.0 (Summa Cum Laude)

Skills

GPU: CUDA/C++, NCCL, Nsight, Energy Profiling, GPU Cluster Management (Slurm, Ansible)
ML & Inference: Speculative Decoding, Quantization (FP8, NVFP4), PyTorch, vLLM, SGLang, FSDP2, Ray
Privacy-Preserving ML: MPC, FHE, FSS