

Krishu Kumar Thapa

Pullman, WA | thapakrishu1997@gmail.com | krishuthapa.github.io | Google Scholar | GitHub

PROFESSIONAL SUMMARY

PhD candidate (5th year) in Computer Science specializing in applied machine learning: reinforcement learning for data center scheduling, supervised forecasting under uncertainty, and efficient LLM training. Backed by 2.5 years of professional software engineering experience with measurable impact: 23.8% reduction in job wait time, 40% memory reduction with 12x faster inference, and forecasting that outperforms existing methods at multi-day and sub-seasonal horizons. Seeking Applied ML Scientist roles focused on translating research into deployed, production-ready systems.

EXPERIENCE

Machine Learning Research Intern

May 2026 – Aug 2026

Hewlett Packard Enterprise (HPE Labs)

Milpitas, CA, USA

- Built RL agents (PPO, DQN) for data center job scheduling using synthetic and real-world (Alibaba) traces
- Cut job wait time by 23.8% versus the strongest baseline, with the lowest SLA violation rate (5.4%)
- Submitted work to AAAI 2027 and the MLForSys workshop at NeurIPS

ML Research Intern

May 2025 – Jul 2025

Argonne National Laboratory

Lemont, IL, USA

- Boosted GPU utilization by 20% optimizing Vision Transformer pretraining with LoRA (ImageNet-1K/21K)
- Achieved 3x throughput and 11% faster runtime scaling PyTorch DDP to 64 GPUs
- Benchmarked distributed training pipelines across node and GPU configurations
- Presented distributed ML infrastructure results to 20+ senior data scientists
- First Author, PreLoRA, AI on HPC Workshop, ISC High Performance

Research Assistant

Aug 2022 – Present

Washington State University

Pullman, WA, USA

- Designing a novel sparse attention mechanism to replace dense self-attention in LLMs (Qwen2/Qwen2.5) for greater compute efficiency (thesis work, in progress)
- Improved snow-water equivalent prediction at 80% of locations with a spatiotemporal attention Transformer (First Author, AAAI 2024)
- Achieved 40% memory reduction and 12x faster inference via hybrid GPU acceleration (Co-author, IEEE TPDS 2025)
- Outperformed existing methods at multi-day and sub-seasonal forecasting with ForeSWE, an uncertainty-aware attention model (First Author, AAAI 2026)
- Mentored 40+ interns and organized ML hackathon with 80+ participants

Software Engineer (L5)

Sep 2019 – Jan 2022

Leapfrog Technology Inc.

Seattle, WA, USA

- Cut R&D workflows by 60% engineering a full-stack drug-discovery web application
- Led 15-member Agile engineering team to deliver two production-ready projects on schedule
- Strengthened team productivity leading developer training, code review, and version control practices

SKILLS

ML Techniques: Supervised Learning, Reinforcement Learning, Time-Series Forecasting, Uncertainty Quantification

ML Architectures: Transformers, Large Language Models (LLMs), ViT, GNNs, RNNs

AI Systems & Acceleration: CUDA, PyTorch DDP, Multi-GPU training, LoRA/PEFT

HPC / Infra: AWS EC2, Distributed training, Profiling, wandb

Programming: Python, C++

Software Engineering: Git/Version Control, Code Review, Agile Development

Data: SQL, Structured/Unstructured Datasets, Statistical Analysis

AI-Assisted Development: Claude Code, GitHub Copilot

AWARDS & HONORS

- WSU Graduate School: Russ and Anne Fuller Interdisciplinary Award (2026)
- WSU EECS: AWS Graduate Student Scholarship (2025)
- WSU EECS: Outstanding Research Assistant (2024)
- AgAID Agathon: Water Intelligence Challenge, 2nd Place (2023)

ACADEMIC SERVICE

- Peer Reviewer, AAAI (2027 Main Track)
- Peer Reviewer, Environmental Data Science (Cambridge University Press)

PUBLICATIONS

- Attention-based models for snow-water equivalent prediction. KK Thapa, B Singh, S Savalkar, A Fern, K Rajagopalan, A Kalyanaraman. Proceedings of the AAAI Conference on Artificial Intelligence, 2024. (First Author)
- HPT: Hybrid acceleration of spatio-temporal attention model training on heterogeneous manycore architectures. S Dahal, P Dhingra, KK Thapa, PP Pande, A Kalyanaraman. IEEE Transactions on Parallel and Distributed Systems, 2025. (Co-author)
- ForeSWE: Forecasting Snow-Water Equivalent with an Uncertainty-Aware Attention Model. KK Thapa, S Savalkar, B Singh, TN Hoang, K Rajagopalan, et al. Proceedings of the AAAI Conference on Artificial Intelligence, 2026. (First Author)
- PreLoRA: Hybrid Pre-training of Vision Transformers with Full Training and Low-Rank Adapters. KK Thapa, R Barik, KT Chitty-Venkata, M Emani, V Vishwanath. AI on HPC Workshop, ISC High Performance Conference. (First Author)

EDUCATION

Washington State University <i>PhD in Computer Science. Thesis/Focus: Sparse Attention Mechanisms for LLMs</i> <i>(Qwen2/Qwen2.5), GPU Optimization. GPA: 4.0.</i> Relevant Coursework: Machine Learning, Neural Networks, Data Science, Advanced Algorithms, Computational Genomics	Pullman, WA <i>Expected: Spring 2027</i>
Kathmandu University <i>B.E. in Computer Engineering. GPA: 3.52.</i>	Dhulikhel, Nepal <i>October 2019</i>