Anurag Mishra

anuragmishra.ofc@gmail.com | (+1) 585-202-5936 | anuragmishra.org | linkedin.com/in/i-anuragmishra | github.com/i-anuragmishra

Education

Master of Science, Artificial Intelligence 2023 – May 2025
Rochester Institute of Technology, Rochester, NY

Bachelor of Technology, Computer Science (Minor: Artificial Intelligence) 2019 – 2023
Sikkim Manipal Institute of Technology, India

Technical Skills

Languages: C/C++, Python, SQL, CUDA | **ML:** Neural Networks, KNN, Transformers, RAG, Probabilistic Modeling
Performance: Profiling, Benchmarking, Regression Debugging, Latency/Throughput Tuning, Memory-Aware Optimization
Parallel: DDP/FSDP, Multicore and Cluster Workloads | **Tools:** PyTorch, JAX, TensorFlow, Linux, Docker, Git, CI/CD, GDB
Systems Exposure: x86/assembly inspection, compiler optimization flags, concurrent computation

Professional Experience

AI Engineer Aug 2025 – Present
Eaton Ventures (Rochester Appliances), Rochester, NY

- Architected and deployed production AI services with tool-calling and retrieval workflows, supporting 10K+ daily requests at 99.9% uptime.
- Reduced end-to-end API latency by 35% by optimizing asynchronous execution, data paths, and model-serving performance.
- Built benchmarking and regression workflows to compare model/service quality and runtime performance across releases.
- Presented performance and reliability results to cross-functional stakeholders to prioritize rollout and optimization workstreams.
- Partnered with product and engineering teams to isolate bottlenecks, debug regressions, and ship reliability improvements.
- Automated validation and deployment checks in CI/CD pipelines, reducing release-cycle time by 60%.

ML Research Assistant, DeFake Project (Multimodal Systems) Oct 2024 – Aug 2025
Rochester Institute of Technology

- Trained multimodal CNN + transformer workloads on 18,000+ videos for lip-sync manipulation detection, improving F1 by 12%.
- Implemented custom C++/CUDA kernels for tensor operations in forward/backward paths, improving training efficiency by 40%.
- Tuned distributed training with DDP/FSDP to improve throughput and memory stability across multi-GPU experiments.
- Generated 50K+ synthetic samples and expanded stress-test coverage, improving out-of-distribution generalization by 18%.
- Built benchmarking workflows across model variants to analyze speed, memory, and quality trade-offs.

ML Research Assistant, LLM Evaluation & Robustness Oct 2024 – Aug 2025
Rochester Institute of Technology, Office of the Provost

- Designed evaluation harnesses for GPT-4, Claude, Llama, and Mistral across reasoning, factuality, and tool-calling reliability tasks.
- Developed retrieval-augmented prompting strategies that reduced hallucination by 35% on factual QA benchmarks.
- Created workload-analysis scripts comparing backend latency and answer quality across model/runtime configurations.
- Built regression test suites to detect behavior shifts early and maintain stable model quality across iteration cycles.
- Authored technical guidance for safer deployment, bias review, and performance monitoring.

ML Research Assistant, Mechanistic Interpretability Jan 2025 – May 2025
Rochester Institute of Technology

- Built activation tracing and profiling tooling in PyTorch/C++ to analyze layer-wise behavior in GPT-style models.
- Identified task-specific circuits in layers 2, 3, and 5; targeted LoRA achieved 40% faster convergence with 75% fewer parameters.
- Developed reproducible benchmarking scripts for convergence, runtime, and quality comparisons across training setups.

Natural Language Processing Intern Nov 2021 – Feb 2022
Textify AI, Remote

- Fine-tuned transformer models for production NLP pipelines and deployed inference APIs with sub-100ms latency.
- Built backend integration workflows for reliable model serving and feature rollout.

Machine Learning Intern Jun 2021 – Aug 2021
Defence Research and Development Organisation (DRDO), New Delhi, India

- Developed probabilistic sensor-fusion and state-estimation models, improving inference speed by 30%.
- Refactored ML prototypes for edge deployment, reducing deployment overhead by 40%.

Selected Publications

Dissecting Chronos: Sparse Autoencoders Reveal Causal Feature Hierarchies in Time Series Foundation Models – OpenReview ICLR 2026
Mechanistic Interpretability of GPT-like Models on Summarization Tasks – arXiv:2505.17073 2025
Automatic Short Answer Grading Using a LSTM Based Approach – IEEE AIC 2023 2023