Anurag Mishra

Curriculum Vitae

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

Research and Engineering Interests

Machine learning systems performance, hardware-aware optimization for neural networks, profiling and regression debugging, mechanistic interpretability, multimodal model training, and reliable deployment of production AI systems.

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 and Algorithms: Neural networks, k-nearest neighbors (KNN), transformers, probabilistic modeling, retrieval-augmented generation (RAG)
Performance Engineering: Benchmarking, profiling, performance-regression triage, latency/throughput optimization, memory-aware tuning
Parallel and Systems: Distributed training (DDP/FSDP), multicore data processing, concurrent computation, Linux, Docker
Frameworks and Tools: PyTorch, JAX, TensorFlow, Git, CI/CD, Weights & Biases
Foundational Exposure: Compiler optimization flags, x86/assembly inspection, native debugging with GDB

Professional Experience

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

- Architected and deployed AI services with retrieval and tool-calling components supporting 10K+ daily requests at 99.9% uptime.
- Reduced end-to-end API latency by 35% through asynchronous orchestration, model-serving optimization, and tighter request-path engineering.
- Built benchmarking and regression-check pipelines to compare model quality and runtime performance across releases.
- Improved deployment cycle time by 60% by implementing CI/CD automation and standardized validation workflows.
- Worked across product and engineering teams to isolate bottlenecks, debug production issues, and ship reliability improvements.

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

- Built and deployed transformer-based text generation APIs for real-time inference with sub-100ms response latency.
- Developed backend services for robust model integration into customer-facing workflows.
- Supported testing and monitoring workflows to improve release stability and operational reliability.

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

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

Research Experience

Machine Learning Research Assistant, DeFake Project Oct 2024 – Aug 2025
Rochester Institute of Technology, Rochester, NY

- Developed multimodal CNN + transformer models for lip-sync manipulation detection on 18,000+ videos, improving F1 by 12%.
- Implemented custom C++/CUDA kernels for tensor operations in forward/backward paths, improving training efficiency by 40%.

- Generated 50K+ synthetic samples and expanded stress-test coverage, improving out-of-distribution generalization by 18%.
- Built evaluation and profiling workflows to compare throughput, memory footprint, and quality across model configurations.

Machine Learning Research Assistant, LLM Evaluation and Reasoning Oct 2024 – Aug 2025
Rochester Institute of Technology, Office of the Provost

- Designed evaluation suites for frontier LLMs (GPT-4, Claude, Llama, Mistral) covering reasoning, reliability, and tool-calling behavior.
- Built synthetic benchmarks and retrieval-augmented prompting pipelines to stress-test factual consistency and model robustness.
- Developed prompting strategies that reduced hallucination by 35% on factual QA benchmarks.
- Authored technical recommendations to translate evaluation findings into practical deployment guidance.

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

- Built activation tracing and profiling tooling in PyTorch/C++ to analyze internal transformer behavior before and after fine-tuning.
- Identified task-specific summarization circuits in layers 2, 3, and 5 with 62% of attention heads showing more focused patterns.
- Applied targeted LoRA interventions that achieved 40% faster convergence (6 vs. 10 epochs) with 75% fewer trainable parameters.
- Awarded Best Poster at the RIT Research Symposium for this work.

Publications

Dissecting Chronos: Sparse Autoencoders Reveal Causal Feature Hierarchies in Time Series Foundation Models ICLR 2026 (Accepted)
First Author. OpenReview

- Accepted at ICLR 2026.

Mechanistic Interpretability of GPT-like Models on Summarization Tasks 2025
First Author. arXiv:2505.17073

- Reverse-engineered summarization behavior in GPT-style models and identified key task-specific internal circuits.

Automatic Short Answer Grading Using a LSTM Based Approach 2023
Co-Author. IEEE AIC 2023

- Developed LSTM-based approach for automated short-answer evaluation in educational settings.

Selected Project

Causal Mechanisms of Backtracking in Reasoning Models 2025
github.com/i-anuragmishra/mats-backtracking

- Studied backtracking behavior in reasoning models and used targeted ablations to identify causally important components.
- Demonstrated large behavioral impact from focused interventions and quantified the relationship between backtracking and answer accuracy.