# Aditya Goyal

+1 608-957-1665 | agoyal33@wisc.edu | linkedin/aditya-goyal | github.com/AdityaGoyal333

## Education

**University of Wisconsin-Madison** Madison, Wi

*Bachelor of Science in Computer Science and Data Science* *Sep. 2023 – May 2027*

**GPA: *3.9/4.0***

**Relevant Coursework**: **High Performance Computing**, Big Data Systems, **ML Research,** Accelerated Honors in Algorithms, Data Structures, **Matrix Methods in ML**, Modeling in R, **Computer Systems, Database Management**

## Technical Skills

Languages: Java, **Python, C, C++**, R, SQL (MySQL, PostgreSQL)
ML: **CUDA, PyTorch, scikit-learn**, GGML, pandas, NumPy, Matplotlib
Backend/Data Engineering: **gRPC**, Flask, FastApi, Redis, Spark
Developer Tools: **Git, GitHub, Docker, Docker Compose**, Postman, Google Cloud, GitHub Copilot

## Experience

**Undergraduate ML Researcher (Sprocket Lab) | UW Madison Computer Sciences** *July 2025 – July 2026*

- Co-developed WARP, a novel weight-space framework for recovering fine-tuning data distributions from base and fine-tuned model weights using gradient signals from simulated training trajectories
- Designed end-to-end WARP experiments across BERT and GPT-2, implemented unsupervised and membership-inference baselines, and **demonstrated 35% and 30% reductions in estimation error**, respectively (MAE: 0.046 and 0.104).
- Proposed future research directions, and **presented findings at the ICML 2026 WSS workshop.**

**Undergraduate ML Researcher (Yin Li's Lab) | UW Madison Computer Sciences** *Nov 2025 – Present*

- Built a reproducible mortality forecasting pipeline for 6K sparse, irregularly sampled NICU recordings.
- Designed a custom Transformer based architecture with cross-attention fusion of seasonal and fine-grained representations, leveraging pretrained Chronos-2 patch embeddings.
- Addressed severe class imbalance using focal loss, prevalence-based bias initialization, and positive-aware batch sampling.
- Improved precision from **0.07 to 0.21 @ 80% recall** across 5-fold cross-validation

**Machine Learning Teaching Assistant | UW Madison Computer Sciences** *Sept 2024 – Present*

- Leading recruitment of future peer mentors and supporting 40+ students every week in CS540 (Intro to AI) and CS 300 (Programming II, Java) on debugging projects and course content, **earning 30+ special mentions** .

**Software Engineering Intern | Frizzle (YC S25)** *May 2026 - Aug 2026*

- Proposed and shipped a rubric-generation feature for IGCSE workflows to prod, achieving **87% eval accuracy** , 0.70$ per rubric while maintaining **130s P95 staging latency**, actively serving **1000+ monthly users**
- Shipped IB and Indian curricula grading workflows at **97% grading accuracy** while maintaining **100s P95 staging latency** after researching formats with 15+ teachers. **The IB feature is set for deployment at 2 schools.**

**Machine Learning Intern | Freecharge** *May 2025 – Aug 2025*

- **Improved spend analysis accuracy by 40%** by engineering AI agents with **AWS Bedrock** to orchestrate tools, generate SQL with RAG context, and deliver charts alongside explanations.
- **Boosted semantic match accuracy by 30%** by building a search engine using **AWS Kendra** over 12,000+ pages and a preprocessing pipeline isolating high-signal fields from noisy webscraped text.
- Enabled real-time branch analytics with YOLO-based video POCs for entrant counting and opening time detection.

**Software Engineering Intern | UW Madison Department of Physics** *Nov 2024 – Feb 2025*

- Implemented Python scripts to train MLPs on mathematical functions and visualize activation flow across layers.
- **Automated large-scale experimentation** by batch processing MLP training runs on randomly generated functions.
- **Improved runtime by 25%** and reduced intermediate memory with vectorized operations and broadcasting.
- Built customizable MLP architectures by parameterizing layers and units, enabling user-specified network designs.

## Projects

**LLM finetuning using LORA** | ***Python, Pytorch*** May 2025

- Fine-tuned FLAN-T5 Large using LoRA for sentiment classification, **reducing trainable parameters by 95%** , achieving 95% accuracy on test set

**ShortNExact** | ***Python****, FastAPI,* ***Docker****, OpenAI SDK, Redis, PgBouncer, HAProxy, Locust* July 2025

- Built a production-grade agentic AI web app that rewrites text to exact word counts while preserving meaning, garnering **2000+ users since deployment**. Containerized services with Docker Compose with health checks and restart policies.
- **Sustained 30+ concurrent users at sub-100 ms median latency** by using **async FastAPI components** and libraries, and validated scalability by load-testing real traffic patterns using **Locust**.
- **Reduced latency by 30% and enabled zero-downtime failover** by deploying **PgBouncer** and **HAProxy**, with secure API access enforced via **Redis**, and API key validation via **PostgreSQL DB** and **SQLAlchemy**.

## Certifications

• Intro to Cloud Computing (IBM) •DL Specialization (Stanford) •ML Specialization (Stanford) •Mathematics for ML (Stanford)