Avinash Bhojanapalli

Rochester NY | +1 585-230-4272 | bhojanapalliavinash@gmail.com | LinkedIn | Github | Google Scholar

Summary

Software & Machine Learning Engineer with a Master's in Computer Science specializing in Data Science. Experienced in bridging the gap between data exploration and production deployment. Proficient in developing scalable backend services (**Go**, **Python**) and optimizing computationally heavy inference pipelines (**Rust**, **PyTorch**). Demonstrated ability to build robust REST APIs, containerize applications with Docker/Kubernetes, and automate workflows to reduce latency and improve system reliability. Passionate about applying core software engineering principles to modern AI challenges.

EDUCATION

Rochester Institute of Technology **Aug 2024 - Apr 2026**
MS Computer Science (Data Science) (GPA: 3.9/4.0) *Rochester, NY*

IIIT Sri City **Dec 2020 - May 2024**
BTech Computer Science and Engineering *India*

SKILLS

- **Languages:** Python, C++, Go, Rust, SQL, Bash
- **Backend & Systems:** RESTful APIs, Microservices, FastAPI, Linux, Git
- **MLOps & Cloud:** Docker, Kubernetes, AWS (SageMaker, Lambda, ECR), CI/CD (GitHub Actions), MLflow, GCP (Vertex AI), Grafana
- **Machine Learning:** PyTorch, TensorFlow, Scikit-learn, Pandas, HuggingFace, Langchain, ONNX
- **Data & Databases:** PostgreSQL, MongoDB, Redis, ChromaDB (Vector DB), Apache Kafka, Neo4j, SQLite

EXPERIENCE

Bonsai Lab RIT | *Machine Learning Engineer* Aug 2025 - Apr 2026

- Architected a modular **Self-Supervised Learning (SSL)** pipeline using **PyTorch**, enabling rapid testing and comparison of model representations and reducing experiment iteration time by **60%**.
- Optimized data loading pipelines with pre-fetching and caching, which increased GPU utilization and reduced training time.
- Implemented automated experiment tracking and artifact versioning, ensuring reproducibility across distributed training runs and standardizing metric reporting for team-wide collaboration.

Incribo | *MLOps Engineer Intern* Jan 2024 - Jul 2024

- Designed and deployed a fault-tolerant microservices architecture for multi-agent AI systems on Kubernetes and Docker, using RabbitMQ for robust asynchronous messaging, which improved system reliability and enabled seamless scaling of AI agents
- Optimized **Large Language Model (LLM)** serving pipelines by implementing **gradient checkpointing** and dynamic batching, resulting in a **15% reduction in GPU memory usage** and improved inference latency.
- Developed a **Go-based** CLI tool to automate container orchestration and deployment workflows, reducing deployment friction and ensuring consistent environments across development and production.
- Integrated **Synthetic Data Generation** workflows using neuroevolution algorithms (CPPN-NEAT), scaling data generation rates by **15%** to support robust model training and validation.

KEY PROJECTS

Crosstem: High-Performance NLP Library | Rust, Python, PyO3, CI/CD | PyPI

- Engineered a multilingual NLP stemming library with a Rust-accelerated core (PyO3) and Python API, delivering 2.7-2.9x speedup over the Python fallback and 15x higher throughput than Porter on benchmark workloads.
- Built a graph-based morphological stemming pipeline (BFS + productivity-threshold scoring) that resolves overstemming and correctly maps across part-of-speech boundaries.
- Shipped and maintained the library on PyPI with parity-tested Rust/Python backends, production packaging, and end-to-end documentation for installation, API usage, and performance validation.

Ledgerline: Vectorless RAG Intelligence Engine | Go, Python, Kafka, Redis | Github

- Engineered an event-driven, vectorless RAG microservices stack, replacing traditional chunking with hierarchical tree navigation to eliminate context loss across complex 200+ page financial filings.
- Architected a high-throughput API Gateway in Go (Fiber) featuring WebSocket streaming for real-time AI reasoning, integrating Redis caching to hot-load document trees and maintain query latency strictly under 3 seconds.
- Orchestrated asynchronous Kafka pipelines for reliable service handoff, integrating an automated LLM-as-a-judge observability system backed by PostgreSQL to continuously evaluate response telemetry and rigorously prevent hallucinations.

Topological Arbitrage Engine: Market-Neutral Quant Strategy | Python, Pandas, SciPy, Flask | Github

- Developed a statistical arbitrage trading engine applying Graph Laplacian diffusion to CAPM residuals, isolating idiosyncratic alpha to exploit structural mean-reversion across a dynamic 80-stock universe.
- Architected a rigorous walk-forward validation pipeline that proved the raw topological signal thrives during market volatility, achieving a 42.5% annualized return and a 1.99 Sharpe ratio on strictly out-of-sample 2025-2026 data.
- Deployed a fully automated live paper-trading infrastructure featuring daily execution, equal-weight position sizing, and state management, integrated with a local web dashboard for real-time portfolio telemetry.

PUBLICATIONS

- A Bhojanapalli, et al..A CALIPSO Observation Based 3-Dimensional Tropospheric Aerosol Classification Model Over the Indian City Delhi.- *In 2023 IEEE International Geoscience and Remote Sensing Symposium, pp. 305-308. IEEE. 16-21 July 2023, Pasadena, CA, USA. DOI:10.1109/IGARSS52108.2023.10282609*
- Z Lyu, et al..Evolving RNNs for Stock Forecasting: A Low Parameter Efficient Alternative to Transformers.*In International Conference on the Applications of Evolutionary Computation (Part of EvoStar), pp. 36-54. Springer, Cham. 17 April 2025, Trieste, Italy. DOI:10.1007/978-3-031-90065-5_3*