Yashdeep Prasad

+1 (929) 919-8891 | yp2693@nyu.edu | Brooklyn, New York, NY, USA | linkedin.com/in/yashdeep18121/ | github.com/prasad-yashdeep

Education

New York University — **August 2024 - May 2026**
Master's, Computer Engineering — *GPA: 3.8*

IIIT, Delhi — **August 2018 - May 2022**
Bachelor's, Computer Science — *GPA: 3.8*

Professional Experience

DeepMind — **Remote**
Applied AI Developer (Google Summer of Code) — *June 2025 - September 2025*
- Redesigned benchmarking protocols for Gemma LLM, boosting function calling accuracy by 30% through advanced Python, C++, and Go optimizations on Linux, leveraging CoT and ReAct for enhanced prompt engineering.
- Drove performance improvements in LLMs by 25% through the deployment of advanced clustering and multi-step reasoning techniques, utilizing deep learning and natural language processing to analyze large datasets for risk detection.
- Automated evaluation processes for LLM function-calling, resulting in a 40% reduction in benchmarking time by developing robust pipelines in Python on Linux and applying data mining and clustering methodologies.
- Developed multi-modal pipelines that integrate computer vision and natural language processing, enhancing agent workflows and achieving a 50% increase in function-calling efficiency in real-time scenarios.

Way.com — **Fremont, CA, USA**
AI Engineering Intern — *June 2025 - August 2025*
- Spearheaded the creation of a custom automated ticketing system using Python and C++, resulting in a 30% increase in workflow efficiency through advanced data mining and clustering techniques in a Linux environment.
- Developed and deployed a deep learning-based chatbot leveraging NLP, improving user interaction response times by 40%, thus enhancing automated support capabilities within the platform.
- Architected and implemented a secure MCP server in Go to expose ticketing APIs, enabling seamless integration of large language models and backend services, thereby boosting support scalability.

Reliance Jio Infocomm Limited — **Bengaluru, KA, India**
AI Engineer — *November 2022 - August 2024*
- Drove a 30% increase in API automation efficiency by leveraging Python and deep learning to fine-tune transformer models (GPT, BERT), enhancing NLP capabilities for internal API calls.
- Engineered a robust orchestration service using Python, Airflow, Kubernetes, and MLflow on Linux, significantly enhancing ML workflow reliability and enabling seamless deployment of computer vision models.
- Implemented advanced clustering algorithms and optimized data structures in Python, reducing data redundancy by 20% and elevating data quality for e-commerce applications.
- Accelerated ML experimentation and deployment cycles by 25% through the development of scalable pipeline components in Python, Go, and C++, streamlining workflows for reinforcement learning and deep learning.

Projects & Outside Experience

NYU Self Driving (VIP Team) — **New York, USA**
ML Algorithm Developer
- Engineered JEPA-based pipelines in Python and C++ for LiDAR-camera fusion, enhancing obstacle anticipation in autonomous driving perception, resulting in a 25% increase in detection accuracy.
- Developed innovative deep learning architectures leveraging data mining and clustering methodologies, achieving a 30% improvement in real-time sensor data processing accuracy through advanced data structures in Python.
- Automated deployment workflows for training and inference on NVIDIA hardware, optimizing performance on Linux platforms and reducing model inference time by 40%.
- *Link to project*

Active Learning and Teaching
Full Stack Engineer
- Optimized backend processing for a cross-platform teaching app with 500+ daily active users, achieving a 50% performance enhancement through GCP, Firebase, and Python, leveraging machine learning for tailored content delivery.
- Enhanced user engagement by 40% by implementing deep learning and NLP techniques in Python to cluster educational content, leading to higher user satisfaction and interaction rates.
- Refactored critical backend modules in a Linux environment using C++, increasing app reliability and scalability, which supported a 30% rise in daily active users.
- Accelerated backend development processes by integrating Go for microservices and Docker, streamlining the deployment of computer vision APIs for automated moderation of user-generated content.
- *Link to project*

Adfame
- Engineered a cutting-edge end-to-end ML-Ops pipeline for an advertising platform, accelerating deployment time by 30% and driving operational efficiency through optimized data ingestion and real-time model monitoring.
- Designed a dynamic training and retraining framework utilizing LoRA for the WAN2.1 text-to-video generation module, leveraging diffusion models to enhance output quality by 25% and processing speed by 40%.
- Integrated MLflow for comprehensive experiment tracking and deployed a Postgres database for real-time metrics storage, boosting model performance visibility and accessibility by 50%.
- Implemented Ray Train and DeepSpeed for sophisticated memory partitioning and pipeline parallelism, increasing training efficiency by 35% and ensuring scalability across cloud environments.
- Collaborated with cross-functional teams to deploy high-quality visuals, utilizing Flask and ONNX to optimize inference services, achieving a 20% improvement in integration speed and user experience.
- *Link to project*

Skills

Skills: Python, Java, JavaScript, React.js, SQL, PowerShell, MongoDB, AWS, Pytorch, Tensorflow, Docker, Machine Learning, Computer Vision, FastAPI, CI/CD, GPU
Languages: French, Hindi