Pranav Mishra

773-280-4615 | pmishr23@uic.edu | LinkedIn/pranavgamedev | Github/PranavMishra17 | Portfolio | Chicago, IL

Education

Master of Science, Computer Science ***Aug 2023 - May 2025***
University of Illinois at Chicago, Illinois, USA | [Graduate Assistant]
Coursework: Advanced ML, Applied AI, Advanced NLP, Computer Vision, VR, Game Design, Object-Oriented Programming

Bachelor of Science, Computer Science and Engineering ***Aug 2019 - June 2023***
Dayananda Sagar College of Engineering, Bangalore, Karnataka, India
Coursework: AI & ML, Software Engineering, Computer Networks, Operating Systems, Data Structures & Algorithms, DBMS

Technical Experience

AI Engineer | *WheelPrice* ↗ | Charlotte, NC ***July 2025 - Mar 2026***

- Built production ML classification system in PyTorch achieving 94% accuracy on 50k+ samples; implemented data preprocessing pipelines and model evaluation frameworks supporting real-time prediction API serving 5k+ requests/day
- Built CMS with microservices architecture on MERN stack handling 10-20k daily users (4× growth); reduced API response time from 1.5s to 300ms through MongoDB query optimization, Redis caching, & connection pooling; with 99.9% uptime

Research Software Engineer | *UIC- VARE Lab* ↗ | Chicago, IL ***Feb 2024 - May 2026****

- Designed RESTful APIs in Python with PostgreSQL, handling 1M+ records via query optimization & auth middleware
- Built audio classification pipeline benchmarking feature-engineered (MFCC+DNN: 98.52%), CNN-based and Transformer approaches on 100k samples; deployed optimized API with 150ms p95 latency via INT8 quantization and batch inference ↗

Research & Publications

TeamMedAgents: SLM-based multi-agent medical reasoning [ICML-2026 Submission] ***GitHub | Paper | SLMs-v2.0***

- Built multi-agent medical reasoning system with Google ADK achieving 77.63% accuracy across 8 benchmarks; Added tool-calling for external retrieval, reasoning traces for decision transparency, and trust-weighted voting for consensus
- Optimized 4B SLMs in multi-agent system; achieved 77.63% accuracy and 3.1× inference speedup vs frontier LLMs

MetaRAG: Metadata Enrichment for RAG Systems [ACCEPTED for IEEE CAI-2026] ***GitHub| IEEE Paper***

- Designed retrieval system achieving 82.5% precision and 0.925 Hit@10 through LLM metadata generation, cross-encoder reranking, and TF-IDF weighted embeddings; reduced hallucinations 25% via automated evaluation framework
- Deployed production LLM system on AWS using SageMaker and MLflow for model versioning and experiment tracking; handled 10k queries/day for code translation and automation tasks; maintained p99 latency <300ms with CI/CD pipelines

Projects

MockFlow-AI: Real-Time Voice Interview Platform with Multi-Agent Architecture ***GitHub | Website***

- Architected real-time voice interview platform with FSM-driven multi-agent orchestration with streaming STT-TTS pipelines
- Achieved sub-400ms end-to-end latency (5× faster than polling-based baseline); optimized WebSocket connections, implemented automated feedback loop, RAG for personalised questions, JWT authentication, and BYOK architecture

SnakeAI-MLOps: Multi-Agent Reinforcement Learning Snake Game ***GitHub | Demo***

- Engineered RL framework with reward shaping, experience replay & hyperparameter optimization in C++ with LibTorch
- Built MLOps pipeline with model versioning, automated testing, Docker CI/CD, and deployment monitoring with gameplay

MedRAG Avatar Platform - Intelligent Healthcare Conversational AI System ***GitHub | Demo***

- Built multi-modal RAG platform with FastAPI backend, Cosmos DB & Azure Speech Services achieving near 100% retrieval accuracy; implemented TTS generation, avatar rendering, and gesture animation with <250ms end-to-end latency

Skills & Extracurricular

TECHNICAL SKILLS: Python, JavaScript, TypeScript, Java, C++, C#, Rust, PyTorch, TensorFlow, NumPy, Pandas, LangChain, SQL, PostgreSQL, MongoDB, GO, Redis, React, Node.js, FastAPI, Flask, Docker, Kubernetes, Git, CI/CD, AWS, Azure, MLflow

APPLICATIONS: AI/ML Engineering, Full-Stack Development, Production ML Systems, LLM Applications, RAG Systems, API Development, Microservices Architecture, System Design, MLOps, CI/CD, Performance Optimization, Cloud Infrastructure

Winner of MIT XR Hackathon | Built Meta Quest 3 app using Unity and Hugging Face for spatial data visualization ↗

INFORMS Analytics+ Speaker | Presented MetaRAG to 700+ professionals | First place HINT 5.0 Web3 museum for NFTs ↗