HARSHIT SONI

New York, NY | (347) 818-5361 | soniharshit2000@gmail.com | Linkedin (harshitsoni01) | harshitsoni.com | GitHub (harshitsoni01)

PROFESSIONAL EXPERIENCE

NextLM Inc. | United States, Remote

Head of Engineering, Founding Team — *Aug 2026 – Present*

- Lead engineering and 0-to-1 product development for NextLM's AI platform, owning full-stack architecture and hands-on coding across React, TypeScript, and Python, scaling to 100+ customers, including enterprise accounts with custom CRM integrations.
- Architect large-scale AI and data infrastructure processing 35B+ rows daily across 350M+ profiles, powering prospecting, enrichment, and buyer-intent intelligence at scale
- Productionize and scale Savant, NextLM's proprietary prospecting model, outperforming frontier models on internal benchmarks and deployed on NVIDIA A100 GPUs with LoRA adapters
- Owned cloud architecture across GCP, APIs, microservices, CI/CD, Docker, and Kubernetes, working directly with Google Cloud and NVIDIA engineers on AI infrastructure and model deployment
- Led engineering efforts to bring NextLM into Gemini Enterprise, advancing the platform's Google-native architecture and enterprise AI distribution

Senior Machine Learning Engineer — *Jan 2026 – Aug 2026*

- Built and deployed production LLM applications, AI agents, and ML systems powering NextLM's sales intelligence platform
- Developed scalable backend services, APIs, data pipelines, and enterprise integrations using TypeScript, Python, and GCP
- Shipped customer-facing AI features from prototype to production while improving platform performance, reliability, and scalability

Cognitive Network Solutions Inc (CNS), *Machine Learning Ops Engineer* | Ridgefield Park, NJ — *Aug 2025 - Dec 2025*

- Developed neural network models for RAN performance prediction and cell tower tilt optimization, enhancing network efficiency
- Designed and deployed end-to-end MLOps workflows for 6G wireless optimization graph RL models, reducing model deployment time and improving reliability across production systems
- Built an automated data pipeline processing 100K+ data every 5 min from Nvidia Sionna using FastAPI, MCP endpoints, & GCP, enabling real-time model training. Containerized and deployed microservices with Docker & Kubernetes, ensuring scalable AI system
- Collaborated with C-suite to align AI-driven cloud systems with CNS's vision for self-optimizing 6G (AI-RAN) networks

Julius Education, *Machine Learning Engineer* | New York, NY — *Jun 2024 – Sept 2024*

- Built ML-powered career dashboards using RAG, prompt engineering, and LLMs to improve renewable-energy career data
- Developed data pipelines, web scraping, and cloud integrations for government datasets and client projects, including NREL

Trust Scout, *Software Engineering Intern* | New York, NY — *Jun 2023 – Aug 2023*

- Built an automated email marketing system that increased client replies by 40% using Python, AWS, Django, SQL, APIs, and AI
- Analyzed 10M real estate transactions and 1.6M agents using ML and statistical modeling to generate business insights

TECHNICAL SKILLS

Languages: Python, TypeScript, SQL, JavaScript, C++, Java, Shell, HTML/CSS
AI/ML: LLMs, AI Agents, RAG, MCP, Machine Learning, Deep Learning, NLP, TensorFlow, PyTorch, Scikit-learn, CUDA
Backend & Data: FastAPI, Flask, REST APIs, Microservices, ETL/Data Pipelines, PostgreSQL, Pandas, NumPy
Cloud & Infrastructure: GCP, AWS (S3, EC2, IAM), Docker, Kubernetes, CI/CD, Git
Engineering: System Architecture, Distributed Systems, Enterprise Integrations, Data Infrastructure, MLOps, Automation

EDUCATION

Binghamton University, State University of New York

Master of Science in Computer Science, Artificial Intelligence Track | **Cumulative GPA:** *3.6 /4.00*

RESEARCH EXPERIENCE

Binghamton University, *Research Assistant – Human-Computer Interaction* — *Jan 2024 – Mar 2024*

- Researched ML model integrating computer vision and natural language processing (NLP) to translate real-time ASL into English text
- Improved hand recognition accuracy using neural networks and image processing for more precise ASL interpretation

Binghamton University, *Research Assistant – Semiconductor Fabrication* — *Jan 2023 – May 2023*

- Researched ML autoencoder models to develop microscopic filter designs for faster semiconductor manufacturing
- Trained an adversarial model (PyTorch, CUDA) to predict and generate filter shapes based on light transmission across wavelengths

IEEE, Published Research Paper Title **'Fake News Detection Using NLP and Logistic Regression'** — *Sept 2021*

PROJECT EXPERIENCE

Zillow Page Summarizer: Chrome Extension Assistant, *Independent Project* (link) — *Jul 2024 – Sept 2024*

- Developed an extension that summarizes Zillow property listings, providing real-time insights using web scraping and NLP
- Built the extension's backend using the Flask framework and integrated an LLM for sentence processing and APIs for data retrieval
- Tech stack: LLM, Python, Flask, Pyzill, API, HTML/CSS, JavaScript, Chrome Extension Framework, Natural Language Processing

LEADERSHIP & HONORS

Hackathon (*Secured 2nd out of the 49 teams)* (link) — *Jul 2020*

- Led a team of three to develop a '**Fake News Detection Model using Machine Learning and Natural Language Processing'**
- Built an ML model to classify news as "fake" or "real," providing users with verified news and enhancing digital media's reliability