

Navya Sree Yellina

✉ navyasreechoudhary@gmail.com 📞 +1 (636)-484-2723 📍 Saint Louis 🌐 navya-sree-yellina

Professional Summary:

Gen AI Engineer with 4+ years developing enterprise-scale generative AI solutions, specializing in deep learning, transformers, and large language models. Delivered 40% performance improvements through machine learning optimization and PyTorch implementation, with proven expertise in ethical AI development, data privacy, and MLOps practices across AWS, GCP, and Azure cloud platforms serving 500+ concurrent users.

Education

Saint Louis University, M.Sc. in Computer Science (Thesis Graduate) Aug 2023 – May 2025

- **Thesis:** Privacy Threats in Continuous Learning (Focused on Machine Learning Security)
- **Relevant Coursework:** Deep Learning, Distributed Systems, Machine Learning, Performance Analysis, Transformers

Koneru Lakshmaiah University, B.Sc. in Computer Science Aug 2017 – May 2021

- **Minor Degree:** Artificial Intelligence
- **Relevant Coursework:** Neural Networks, NLP, Data Structures, Algorithms, Database Systems, Ethics in AI

Professional Experience

Gemini Consulting & Services – Chesterfield, MO Jan 2025 – Present
Generative AI Engineer Intern

- Architected enterprise generative AI platform using OpenAI GPT API, transformers, and deep learning models with PyTorch and TensorFlow, reducing information retrieval latency by 40% (2.1s → 1.26s) while supporting 500+ concurrent users with 90% system uptime improvement
- Implemented RAG framework with Python, LangChain, and machine learning optimization techniques, achieving 25% improvement in NLP accuracy across 10,000+ production queries through systematic transformers fine-tuning and precision enhancement
- Deployed multi-channel AI agents using Python, Azure APIs, and MLOps best practices for contact center operations, increasing response throughput 30% (450→585 requests/min) with focus on ethical AI principles and data privacy compliance
- Established automated MLOps pipeline using Docker, Kubernetes, AWS SageMaker, and Git workflows, accelerating deep learning model deployment cycles by 35% while reducing manual errors from 15% to 3% through CI/CD automation
- Applied computer science principles and software engineering best practices to ensure scalable, reliable AI infrastructure with comprehensive unit testing and performance monitoring across cloud platforms

Environment: Python, PyTorch, TensorFlow, OpenAI GPT API, LangChain, FastAPI, Docker, Kubernetes, AWS SageMaker, Azure APIs, Git, GitHub Actions, Supabase, Microsoft Graph API

Oracle Cerner – Bengaluru, IND May 2021 - July 2023
Systems Engineer

- Built distributed machine learning monitoring system using Python and deep learning frameworks for 50+ microservices, reducing incident response time by 20% (45→36 minutes) while maintaining 99.9% uptime across 2.5M+ daily transactions
- Developed high-performance ETL pipelines with Python, SQL, and TensorFlow for Oracle-to-PostgreSQL migration, improving query performance by 25% and reducing hosting costs by \$50K annually through optimized data processing algorithms
- Automated cloud infrastructure provisioning using Python, Terraform, Docker, and Kubernetes across AWS and Azure platforms, managing 200+ S3 buckets and 50+ EC2 instances while implementing MLOps practices for containerized services
- Implemented Git-based CI/CD deployment workflows with automated validation and precision testing, reducing high-risk production incidents by 30% through systematic process auditing and machine learning monitoring

- Provided technical guidance and mentorship to 2 junior developers on software design patterns, unit testing frameworks, and engineering principles while fostering collaboration skills across cross-functional teams.

Environment: Python, Zabbix, TensorFlow, Oracle Database, PostgreSQL, Terraform, CloudFormation, Docker, Kubernetes, AWS (S3, EC2), Azure, Git, CI/CD Pipelines, Linux, Shell Scripting

Televerge Communications – Bengaluru, IND

Jan 2021 - April 2021

Software Engineer Intern

- Optimized backend systems using Java, Python, and machine learning algorithms, scaling from 7K to 10K+ daily API requests with a 30% throughput improvement and a 15% memory reduction through computer science optimization techniques
- Developed REST API integrations with React frontend and SQL databases, improving data delivery speed by 40% (500ms→300ms response time) while implementing data privacy measures for 5,000+ active users
- Created reusable software libraries for network protocol implementation using Git version control, increasing development efficiency by 25% through modular design and comprehensive testing frameworks
- Processed 1M+ network events daily using distributed computing and machine learning processing, demonstrating research capabilities and precision in an agile development environment

Environment: Java, Spring Boot, MongoDB, React, JavaScript, Python, REST APIs, Git, Maven, Linux, Network Protocols, SQL, Unit Testing Frameworks

Technical Skills

Generative AI & Deep Learning: Transformers (GPT, BERT, T5), Large Language Models, OpenAI API, LangChain, RAG Frameworks, Prompt Engineering, Model Fine-tuning, Weights & Biases, Precision Optimization

Machine Learning Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, Keras, MLflow, Computer Vision, NLP


Cloud & MLOps: AWS (SageMaker, Lambda, EC2, S3), Azure ML Studio, GCP AI Platform, Docker, Kubernetes, CI/CD Pipelines, GitHub Actions

Programming & Development: Python, SQL, JavaScript, Java, Git, FastAPI, Flask, React, RESTful APIs, Microservices Architecture

Data & Infrastructure: PostgreSQL, MongoDB, ETL Pipelines, Pandas, NumPy, Data Privacy Compliance, Distributed Systems

Research & Ethics: Ethical AI Development, Model Interpretability, Privacy-Preserving ML, Research Publications, Computer Science Theory

Publications and Achievements

Publication: “Inspecting CNN and ANN Algorithms using Digit Recognition Model,” IRJET, Jun 2020 

Current Research: Privacy-preserving techniques in continuous learning environments, focusing on differential privacy and deep learning approaches. Active R&D in emerging technologies including machine learning security, computer vision applications

Recognition: Women Entrepreneur of the Year (2018) for driving business innovation and growth.

Award: Employee of the Month for reducing high-risk incidents by 30% through process auditing