

# Navya Sree Yellina

📍 Saint Louis | ✉️ navyasreechoudhary@gmail.com | 📞 (636)484-2723

in <https://www.linkedin.com/in/navya-sree-yellina/> | 🌐 <https://github.com/Navya-sree-yellina>

## Professional Summary:

---

Generative AI Engineer with 4+ years of experience developing and deploying enterprise-grade AI solutions across healthcare, government, and consulting domains. Proven expertise in architecting intelligent systems using large language models, RAG frameworks, and transformer architectures, delivering measurable impact including 40% reduction in information retrieval latency, 25% improvement in NLP accuracy, and 90% system uptime enhancement. Specialized in full-stack AI development—from solution design and prompt engineering to MLOps deployment and scaling multi-channel AI agents for 500+ concurrent users. Expert in leveraging both cutting-edge technologies (OpenAI GPT API, LangChain, PyTorch, TensorFlow) and enterprise platforms (AWS SageMaker, Azure ML Studio, Microsoft Power Platform) to translate complex business requirements into scalable, secure, and privacy-compliant AI solutions that drive organizational transformation and measurable ROI.

## 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, Statistics, Data Analysis, 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 AI platform using OpenAI GPT API, Llama LLM, LlamaIndex, LangChain, and LangGraph with FastAPI to address slow information retrieval (2.1s latency) and scalability issues, reducing latency by 40% (2.1s to 1.26s) while supporting 500+ concurrent users with 90% system uptime improvement for enhanced customer experience
- Developed specialized AI agents including Email Agent for intelligent email classification and automated responses, WhatsApp Agent for real-time customer support automation, and Notifications Agent for cross-platform context-aware communications, delivering 30% increase in response throughput (450→585 requests/min) with 20% latency reduction and improved customer satisfaction
- Implemented RAG framework using Python, Supabase embeddings, and LangChain with extensive research in generative AI technologies to improve contextual accuracy of AI responses across 10,000+ production queries, achieving 25% improvement in contextual accuracy through systematic machine learning optimization and natural language processing (NLP)
- Deployed AI models into production environments using Docker, TensorFlow Serving, and Azure Servers while developing APIs with FastAPI and Python to ensure seamless backend-frontend integration, accelerating model deployment cycles by 35% and reducing manual errors from 15% to 3% following DevOps best practices
- Integrated AI Agent functionalities into dynamic user interfaces using Node.js, React.js, and Next.js while collaborating with front-end developers to align technical functionality with user-centric design principles, delivering enhanced user experience across web applications serving enterprise clients
- Conducted comprehensive manual and automated testing, quality assurance (QA), and debugging activities while utilizing GitHub for code management and version control, significantly enhancing robustness and reliability of AI solutions with industry-standard practices and comprehensive documentation
- Successfully collaborated with international offshore team members across different time zones while actively participating in agile methodologies including sprint planning, stand-up meetings, and retrospectives,

demonstrating effective cross-cultural communication and project management skills in global development environment

**Environment:** Python, FastAPI, OpenAI GPT API, Llama LLM, LlamaIndex, LangChain, LangGraph, Docker, Azure, Node.js, React.js, Next.js, GitHub, Supabase

**Oracle Cerner** – Bengaluru, IND  
**Systems Engineer**

May 2021 - July 2023

- Designed distributed monitoring system using Python and Zabbix for 50+ microservices within Cerner Millennium solution architecture to reduce incident response times, achieving 20% improvement (45→36 minutes) while maintaining 99.9% uptime across 2.5M+ daily transactions through automated monitoring and alert handling protocols
- Built high-performance ETL pipelines for Oracle-to-PostgreSQL migration across multiple client regions to optimize data processing workflows, improving query performance by 25% and reducing hosting costs by \$50K annually through systematic database optimization and migration strategies supporting enterprise healthcare solutions
- Automated cloud infrastructure provisioning using Python, Terraform, and CloudFormation for AWS environments to streamline deployment processes across multiple domains and environments, managing 200+ S3 buckets and 50+ EC2 instances while containerizing 15+ services with Docker and Kubernetes following DevOps methodologies
- Implemented Git-based CI/CD pipelines with automated validation scripts to enhance production domain activities including refreshes, migrations, package installations, and security patching, reducing high-risk production incidents by 30% through systematic process auditing and continuous integration practices
- Performed on-call rotation support every 4-6 weeks ensuring high availability of healthcare systems while troubleshooting production issues using Shell Scripting, BMC Patrol, and custom monitoring solutions, maintaining standard operating procedures and knowledge base documentation for 24/7 system reliability
- Provided technical mentorship to 2+ junior developers on system troubleshooting, database migration best practices, and enterprise architecture principles while streamlining support processes to improve communication between technical teams and end-users across global healthcare operations

**Environment:** Python, Zabbix, Shell Scripting, BMC Patrol, Oracle Database, PostgreSQL, AWS (EC2, S3), Docker, Kubernetes, Terraform, CloudFormation, Git, Cerner Millennium

**Televerge Communications** – Bengaluru, IND  
**Software Engineer Intern**

Jan 2021 - April 2021

- Completed comprehensive web development training using JavaScript, HTML, and CSS to build foundational programming skills, progressing from conceptual learning to hands-on implementation of minor web projects
- Advanced to Python-based development projects focusing on automation systems, implementing advanced Python modules and SQL programming while applying problem-solving and practical skill application principles
- Optimized backend automation systems using Java, Spring Boot, and MongoDB while serving as testing specialist, scaling from 7K to 10K+ daily API requests with 30% throughput improvement and 15% memory reduction
- Developed REST API integrations with React frontend using JavaScript to create responsive web applications, improving data delivery speed by 40% (500ms→300ms response time) for 5,000+ active users
- Built reusable software libraries for automation testing and network protocol implementation following software design principles, improving development efficiency by 25% through modular design and systematic testing frameworks

**Environment:** JavaScript, HTML, CSS, Python, SQL, Java, Spring Boot, MongoDB, React, REST APIs, Network Protocols, Git, Automation Testing

**National Informatics Center** – Hyderabad, IND  
**Web Developer Intern**

July 2020 - Dec 2020

- Developed comprehensive PDF converter web application for Government of India to digitize document processing workflows, implementing secure file upload, conversion, and download functionalities that reduced manual processing time by 60%

- Architected scalable backend system using Node.js and Express.js with MongoDB database to handle high-volume document conversion requests, supporting 1000+ concurrent users and processing 5000+ daily PDF conversions with 99.5% uptime
- Implemented advanced PDF manipulation features including document merging, splitting, watermarking, and format conversion using PDF-lib and custom algorithms, enabling efficient management of official documents with standardized formatting
- Designed responsive frontend interface using HTML5, CSS3, Bootstrap, and JavaScript to ensure accessibility across devices, creating intuitive user experience that reduced employee training time by 40% while maintaining web accessibility compliance
- Integrated secure file handling protocols with encryption and access controls to protect sensitive documents, implementing role-based authentication, audit trails, and data retention policies that met government cybersecurity requirements
- Collaborated with IT administrators to gather requirements, conduct user acceptance testing, and deploy across 15+ departments, providing technical documentation and training materials for successful adoption by 500+ employees

**Environment:** JavaScript, HTML5, CSS3, Node.js, Express.js, MongoDB, PDF-lib, Multer, Bootstrap, Git, Government Standards, Security Protocols

## Technical Skills

---

**Generative AI & Deep Learning:** Transformers (GPT, BERT, T5, LLAMA, Claude), Large Language Models (GPT-4, GPT-4o, Claude-3, LLAMA 2/3, Mistral), OpenAI API, Anthropic API, LangChain, LangGraph, RAG Frameworks, Prompt Engineering, Advanced Prompt Techniques (Chain-of-Thought, Few-shot), Model Fine-tuning, LoRA/QLoRA, PEFT, Reinforcement Learning from Human Feedback (RLHF), AI/ML, Hugging Face Transformers, Agent-based Systems

**Vector Databases & Embeddings:** Vector Databases (Pinecone, ChromaDB, Weaviate), Vector Embeddings, Semantic Search, Similarity Search, Embedding Models (OpenAI Embeddings, Sentence-BERT), Vector Indexing, High-dimensional Data Processing

**AI/ML Platform & Tooling:** AWS (SageMaker, Lambda, EC2, S3, Bedrock), Azure ML Studio, Azure OpenAI Service, Databricks, GCP AI Platform, Vertex AI, Docker, Kubernetes, CI/CD Pipelines, GitHub Actions, MLflow, Weights & Biases, PyTorch, TensorFlow, scikit-learn, Ollama, LM Studio

**Low-Code/Pro-Code:** Microsoft Power Platform, Copilot Studio, Microsoft Copilot, RPA, Zapier, AutoGen, LangSmith, Streamlit, Gradio

**AI Ethics & Safety:** AI Ethics, Responsible AI, Bias Detection & Mitigation, AI Governance, Model Interpretability, Explainable AI (XAI), AI Safety, Content Filtering, Guardrails


**Programming & Development:** Python, SQL, JavaScript, Java, TypeScript, Go, Git, FastAPI, Flask, Streamlit, React, Next.js, RESTful APIs, GraphQL, Microservices Architecture, Automated Testing, Unit Testing Frameworks, API Development & Integration Architecture, Automated Testing, Unit Testing Frameworks

**Data & Infrastructure:** PostgreSQL, MongoDB, Redis, Elasticsearch, ETL Pipelines, Pandas, NumPy, Data Privacy Compliance, GDPR Compliance, Distributed Systems, Data Pipeline, Feature Engineering, Data Preprocessing, Unstructured Data Processing

**Soft Skills & Core Competencies:** Technical Expertise, Cross-functional Collaboration, Stakeholder Communication, Adaptability, Complex Problem-Solving, Technical Leadership, Mentorship, Solution Architecture, Business Acumen, Articulate Technical Concepts to Non-Technical Audiences

## Publications and Recognition

---

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

**Current Research:** Privacy-preserving techniques in continuous learning environments with a focus on differential privacy applications in fraud detection systems and patient safety analytics using artificial intelligence and advanced statistical methods

**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

## Research Experience and Competitions

---

### **Tata Institute of Fundamental Research (TIFR) – Hyd, IND**

Jun 2019- May 2020

- Selected for a prestigious research program among competitive applicants nationwide
- Conducted advanced research in computational sciences and theoretical foundations
- Collaborated with leading researchers on cutting-edge projects in mathematics and computer science

### **Hackathon Competitions**

- **STLHack** - Participated in competitive coding and innovation challenge
- **SLUHack** - Competed in Saint Louis University's premier hackathon event