KAVYA SREE CHANDHI

Mail id: Kavyasreechandhi@gmail.com | **Contact Number:** 9406298898 | **Location:** Herriman, Utah
LinkedIn: https://www.linkedin.com/in/chandhi-kavya-sree/ |
GitHub: https://github.com/kavya-sree-chandhi |
Portfolio:

PROFESSIONAL SUMMARY

- Generative AI Engineer with **3+ years of experience** in **Artificial Intelligence, Machine Learning, and Data Engineering**, including **1 year of expertise in Generative AI** with **Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), fine-tuning, and cloud-based deployments**.
- Designed and implemented AI-powered solutions such as an **Attendance Monitoring System** with **95%+ accuracy**, **dynamic chatbots**, and **real-time analytics platforms** across **healthcare, finance, and retail domains**.
- Skilled in **Python programming, API development, and automation** with strong expertise in **Natural Language Processing (NLP), Computer Vision, and Deep Learning** for tasks such as **classification, summarization, and object detection**.
- Proficient in **Frontend Development** using **Python, HTML, CSS, and JavaScript**; completed a volunteer internship at **NTARI (Network Theory Applied Research Institute)** as a **Frontend Developer** contributing to **AI-driven web applications**.
- Strong experience in **data visualization and analytics** using **Power BI, Tableau, Grafana, Matplotlib, and Seaborn** to generate insights and support decision-making.
- Adept in **MLOps and DevOps practices** including **Git, Docker, Kubernetes, CI/CD pipelines, Apache Kafka, and Airflow** for scalable deployment and monitoring of AI/ML solutions.
- Experienced in **cloud platforms** including **AWS (SageMaker, Bedrock, Lambda, EC2, S3, Redshift)**, **Microsoft Azure (Prompt Flow, Cosmos DB, Bot Framework)**, and **Google Cloud (Vertex AI, AutoML, BigQuery ML, Pub/Sub)**.
- Passionate about building **cloud-native, production-ready Generative AI and Machine Learning solutions** that drive **automation, innovation, and business impact**.

TECHNICAL SKILLS:

- **Programming Languages:** Python, C, JavaScript, HTML, CSS, Linux, TypeScript, SQL, Shell Scripting (Bash)
- **Generative AI:** Large Language Models (LLMs), Transformers, Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-Tuning, AI Agents, Vector Databases (Milvus, ChromaDB), Frameworks (LangChain, Hugging Face, Llama, Groq, OpenAI, Ollama), DALL·E.
- **Machine Learning & Data Science:** Supervised and Unsupervised Learning, Feature Engineering, Predictive Modeling, Cross-Validation, Hyperparameter Tuning, Data Analytics, Statistical Analysis, Scikit-learn
- **Deep Learning & Computer Vision:** Neural Networks (ANN, CNN, RNN, LSTM), Transfer Learning, ResNet, MobileNet, Image Classification, Object Detection, OpenCV, PyTorch, TensorFlow, Keras
- **Natural Language Processing (NLP):** Sentiment Analysis, Emotion Detection, Text Classification, Summarization, Named Entity Recognition (NER), spaCy, NLTK
- **Data Visualization & Analytics:** Power BI, Tableau, Grafana, Matplotlib, Seaborn, Excel, Exploratory Data Analysis (EDA), Predictive Analytics
- **Cloud Platforms:**
  AWS: Bedrock, SageMaker, Lambda, EC2, S3, CloudFront, CloudWatch, API Gateway, Redshift, Kinesis, EMR, QuickSight
  Azure: Azure Prompt Flow, Azure Bot Framework, Azure Cosmos DB
- **Databases:** SQL, MySQL, PostgreSQL, MongoDB, DynamoDB, NoSQL
- **API & Web Development:** RESTful APIs, API Integration, FastAPI, Flask, Streamlit, Django, Postman
- **MLOps & DevOps:** Git, GitHub, Docker, Kubernetes, CI/CD (AWS CodePipeline, CodeBuild), Apache Kafka
- **Tools & IDEs:** Jupyter Notebook, Google Colab, VS Code, PyCharm, Anaconda, Elasticsearch

WORK EXPERIENCE:

~~Client - UnitedHealthcare / Minnetonka, MN~~ **~~Jul 2024 – Present~~**

Role - Generative AI Engineer

Project Description: Developed a **Generative AI-based Contract Intelligence Platform** using **GPT-4, LangChain, and RAG pipelines** to summarize, analyze, and compare legal and compliance documents. Reduced contract review time by **45%** and improved compliance accuracy by **30%**.

Key Responsibilities:

- Designed and implemented a **contract summarization assistant** using **GPT-4 and Hugging Face transformers**, enabling legal teams to extract key provisions faster and improving overall review efficiency.
- Built a **Retrieval-Augmented Generation (RAG) pipeline** with **Pinecone and FAISS**, enabling semantic search across millions of legal documents and improving retrieval precision by **25%**.
- Fine-tuned **GPT-3.5/4 models** on compliance and regulatory datasets, reducing context errors and increasing domain-specific accuracy.
- Developed **clause comparison pipelines** to automatically flag risky or missing terms, reducing manual review time by **20%** and strengthening compliance assurance.
- Integrated **Azure Cosmos DB** for secure document storage and retrieval, ensuring scalability and regulatory compliance.
- Deployed the platform on **Azure Kubernetes Service (AKS)** using **Dockerized microservices**, achieving **99.9% uptime** and enterprise scalability.
- Applied **prompt engineering** strategies along with **AI guardrails and bias detection**, standardizing outputs and ensuring legally compliant responses.
- Built interactive **compliance dashboards** in **Power BI and Streamlit**, giving stakeholders real-time visibility into contract risks, approval timelines, and compliance metrics.
- Integrated with **DocuSign APIs** to streamline the contract lifecycle, cutting administrative delays by **30%**.
- Conducted **A/B testing** with in-house legal teams, reducing review turnaround time by **45%** compared to manual methods.
- Collaborated with a **multidisciplinary team of consultants, lawyers, and engineers**, ensuring outputs aligned with legal frameworks and enabling enterprise-wide adoption.

~~Client - USAA (United Services Automobile Association), USA | TCS, India~~ **~~Aug 2021 – Nov 2022~~**

Role - AI/ML Engineer

Project Description: Built an **AI-driven** Demand Forecasting and Inventory Optimization Platform to predict product demand and optimize stock levels across 500+ retail stores. Leveraged **AWS, PyTorch, and Spark** to reduce stockouts by 19% and cut excess inventory holding costs by 12%. A real-time fraud detection system using Kafka, Spark Streaming, and AWS, boosting fraud.

Key Responsibilities:

- Designed and implemented **time-series forecasting models** using **LSTMs and ARIMA** across 500+ SKUs, improving demand prediction accuracy by **22%** and aligning stock levels with real-world sales trends.
- Processed **multi-terabyte sales and transactional datasets** with **Apache Spark on AWS EMR**, enabling near real-time forecasting and significantly reducing data processing latency.
- Automated **ETL pipelines** using **Apache Airflow, Amazon Redshift, and AWS S3**, cutting manual intervention by **40%** and ensuring continuous, reliable data availability for machine learning workflows.

- Conducted **feature engineering** by incorporating seasonal, regional, and promotional data patterns, improving the robustness and generalizability of forecasting models.
- Deployed forecasting models as **REST APIs on AWS SageMaker**, fully integrated into Walmart's enterprise supply chain system for live operational usage.
- Developed a **store-level replenishment recommendation engine**, reducing stockouts by **19%** and boosting customer satisfaction rates.
- Built real-time **dashboards in Tableau and Grafana**, providing supply chain managers with actionable KPIs such as demand forecasts, reorder cycles, and risk of stockouts.
- Designed and implemented **inventory optimization algorithms**, lowering warehouse holding costs by **12%** while maintaining demand-supply balance.
- Applied **unsupervised anomaly detection models** to flag irregular sales spikes during promotions and seasonal campaigns, enabling faster decision-making.
- Containerized forecasting services using **Docker** and deployed them on **Kubernetes clusters**, ensuring scalable and fault-tolerant operations in cloud environments.
- Authored comprehensive **technical documentation and training guides**, reducing onboarding time for supply chain teams and enabling smoother adoption of AI-driven forecasting systems.

VOLUNTEERING EXPERIENCE:

Client - NTARI (Network Theory Applied Research Institute) **Jun 2025 – Aug 2025**

Role - Frontend Developer

Project Description: Contributed as a volunteer developer to NTARI's **LogicLingo application**, focusing on **frontend architecture, user interface design, and backend system analysis**. Actively participated in **Agile operations via Slack**, ensuring technical delivery and team collaboration.

Key Responsibilities:

- Designed and implemented **responsive user interfaces** using **HTML, CSS, JavaScript, and Python frameworks**, improving overall usability and boosting user satisfaction scores by **25%**.
- Contributed to **frontend architecture** by developing **reusable UI components and layouts**, which enhanced scalability and reduced redundant code by **30%**.
- Developed **interactive features and workflows** that improved navigation efficiency, leading to a **20% faster task completion rate** for end-users.
- Conducted **backend file reviews and system architecture analysis**, identifying integration issues early and reducing bug rates in deployments by **15%**.
- Participated in **Agile ceremonies and Slack-based collaboration**, ensuring **100% on-time delivery** of weekly sprint goals.
- Optimized **frontend performance** through code refactoring and asset optimization, reducing page load times by **35%** across supported devices.
- Enhanced **application accessibility** by implementing **WCAG 2.1 standards**, expanding usability for all users and ensuring compliance with accessibility guidelines.
- Integrated **backend APIs** into frontend modules for real-time, data-driven features, increasing dynamic content delivery efficiency by **40%**.
- Created and maintained **technical documentation** for workflows, improving team knowledge transfer and reducing onboarding time for new developers by **20%**.
- Recognized by project leadership for **initiative and consistency**, demonstrating reliability with a **20-hour weekly contribution schedule** and helping drive the LogicLingo application to critical milestones.

PROJECTS:

AI Research Agent for Healthcare Diagnostics

- Developed an **AI-powered research assistant** using **Python, Streamlit, LangChain, and LLaMA3**, integrating web, academic, and local sources through a **RAG pipeline with FAISS and PyMuPDF** to generate structured healthcare diagnostics reports with citations and automated export in multiple formats.

Career Navigation AI

- Developed an AI-powered career copilot using **LangChain + GPT-4** with a **Streamlit** UI that builds personalized AI/ML learning paths, aggregates resources, and improves career readiness by 40%.

Automatic Attendance Monitoring System(Face Recognition)

- Built an **AI-driven attendance system** using **Python, OpenCV, dlib, and Django** with **ML classifiers (SVM, KNN)**, achieving automated face recognition, secure reporting, and visual analytics that eliminated proxy attendance and improved tracking efficiency.

Image Classification for Disease Detection

- Developed **deep learning models (CNNs, VGG16, ResNet50)** for medical image classification on Brain MRI and Chest X-ray datasets, achieving **98% accuracy** in brain tumor detection with VGG16, supported by preprocessing pipelines using **OpenCV, NumPy, and Pandas** across **10k+ images (~2.7 GB)**.

Youth Smoking and Drug: Analysis & ML Insights

- Analyzed a dataset of **8,993 records** using **Python, Pandas, NumPy, and Seaborn**, performing EDA, hypothesis testing, and ML modeling that improved high-risk youth identification by **18%**, showing **peer influence raised smoking prevalence by 20%** and **school-based interventions reduced substance use by 12%**.

Securing Data Using Blockchain and Artificial Intelligence (AI)

- **Developed SecNet architecture** integrating **Blockchain-based trusted data exchange, AI-powered secure computing, and Django (Python) prototype** for healthcare, enabling patient-controlled hospital data access, tamper-proof recordkeeping, and reducing unauthorized exposure to improve trust in medical record sharing.

CERTIFICATES:

- **Google IT Support Fundamentals** – Coursera/Google
- **Website Development**– 1Stop
- **Oracle Cloud Infrastructure Generative AI Professional** – Oracle
- **AWS Machine Learning Foundations** – AWS & Udacity
- **Fundamentals of Generative AI** – Microsoft

ACADEMICS:

Master of Science in computer science **GPA: 3.916/4**

University of North Texas – Denton, Texas

- Specialized in **Artificial Intelligence, Machine Learning, and Software Engineering**, with hands-on projects in **Generative AI, NLP, and Deep Learning**.
- Applied **Python, Java, TensorFlow, LangChain, and Database Management** to build AI-driven solutions and intelligent applications.
- Completed academic and hackathon projects using **Agile methodologies**, deploying scalable applications.