THARUNESWAR DODDI

+91 6281162655 • tharuneswardoddi@gmail.com • linkedin.com/in/tharuneswar-doddi • https://github.com/Tharuneswar2 • https://tharuneswar2.github.io/

AI/ML Engineer | AI Full Stack Engineer | Intelligent Systems Architect

- AI Full Stack Engineer with 1.5+ years of experience designing and deploying production-grade AI/ML systems and scalable backend architectures. Strong expertise in Python, PyTorch, TensorFlow, Scikit-learn, FastAPI, MongoDB, Qdrant, and AWS. Proven experience building transformer-based LLMs, Retrieval-Augmented Generation (RAG) pipelines, computer vision systems, and real-time distributed backend services.
- Experienced in data preprocessing, machine learning model deployment, REST API engineering, performance optimization, and cloud deployment, delivering scalable AI solutions that improve automation, decision-making, and operational efficiency.

WORK EXPERIENCE

Criativo Software Solutions Pvt Ltd
AI Full Stack Engineer (Python Backend & AI Systems)

- Designed and deployed scalable AI-driven backend systems using FastAPI and Python.
- Built machine learning inference APIs supporting high-concurrency production workloads.
- Improved API response performance by 30% through query optimization and async processing.
- Developed RESTful services for AI pipelines including embedding generation, similarity search, and evaluation engines.
- Performed model integration, backend orchestration, and deployment automation on AWS.
- Conducted API testing and validation using Postman and Requestly, ensuring reliability across production services.
- Participated in Agile development cycles, code reviews, and architecture design discussions.

ULearn
AI & Machine Learning Intern

- Implemented machine learning models using Scikit-learn, TensorFlow, and Python.
- Performed dataset preprocessing, feature engineering, and evaluation metrics analysis.
- Built supervised learning pipelines for classification and regression problems.
- Achieved A+ performance rating for delivering optimized ML implementations.

EDUCATION

MCA
Aditya PG College 01/2025 - Present

MCsAiR
Aditya Degree College 01/2022 - 01/2025

Intermediate (MPC)
P.R. Govt Junior College

SSC
M.G.M.C High School

CERTIFICATIONS

Python Essentials
Cisco

Competitive Coding

Criativo E-Learning

Ethical Hacking

Infosys Springboard

Building RAG Agents with LLMs

NVIDIA

Python, ML & Data Analysis Workshop

AWARDS & SCHOLARSHIPS

1st Place – Robo Race 2K23

PROJECTS

GPT-Style Transformer Model From Scratch

- Designed and trained transformer-based LLM using PyTorch.
- Implemented custom tokenizer, instruction fine-tuning pipeline, and checkpointing.
- Reduced GPU memory consumption by ~35% using mixed-precision training (AMP).
- Built scalable training loop architecture for experimentation.

Face Recognition & Clustering Platform (CampusVibe)

- Built production-ready face recognition system using FastAPI, MongoDB, Qdrant, DeepFace, and Docker.
- Implemented embedding pipelines and vector similarity search supporting large-scale image datasets.
- Designed clustering pipeline to identify visually similar identities automatically.
- Enabled bulk ingestion (ZIP uploads) and high-concurrency processing APIs.

Financial Retrieval-Augmented Generation (RAG) Agent

- Designed semantic search architecture integrating MongoDB and Qdrant.
- Built document ingestion pipeline and embedding indexing workflows.
- Enabled AI-driven financial query answering with contextual reasoning.

Structured OCR Invoice Extraction System

- Implemented structured OCR pipeline using transformer-based document extraction models.
- Optimized image preprocessing to reduce inference memory consumption.
- Integrated structured JSON output pipeline for downstream analytics systems.

ML Predictive Analytics Pipeline (Added for JD Alignment)

- Developed predictive analytics system using Scikit-learn for classification and regression tasks.
- Implemented automated preprocessing, feature engineering, and model selection workflows.
- Evaluated models using Precision, Recall, F1-Score, ROC-AUC, improving prediction accuracy by ~18%.
- Deployed trained models as REST APIs using FastAPI for real-time inference.

Real-Time Chat & Event Streaming Platform

- Developed WebSocket-based real-time messaging backend with MongoDB change streams.
- Designed MQTT-based scalable event streaming architecture for distributed systems.

SKILLS

Programming: C, Java, JavaScript, Python, Shell

Machine Learning: PyTorch, Scikit-learn, TensorFlow, Transformers

AI Systems: Embeddings, GPT-style Models, RAG, Semantic Search

Data Processing: Data Preprocessing, Feature Engineering, NumPy, Pandas

Backend Engineering: FastAPI, MQTT, REST APIs, WebSockets

Databases: MongoDB, Qdrant, SQL

Cloud & DevOps: AWS EC2, Docker, Kubernetes (Fundamentals), Linux, Nginx

Testing & Tools: API Testing, Git, Postman, Requestly

Software Engineering: Agile Development, Algorithms, Data Structures, OOP

Computer Vision: DeepFace, OpenCV

Automation: n8n, UiPath