

Deshant Singh Bani

✉ deshantbani@gmail.com

☎ (+91) 9354532389

🌐 linkedin.com/in/deshantbani

EXPERIENCE

Jio Platforms Limited.....
AI Intern Dec. 2024 - Jan. 2025, Noida

- Engineered a robust multi-agent framework, enabling seamless collaboration among agents and achieving a 100% task completion rate.
- Designed modifications to support smaller LLMs (e.g., LLaMA 3.2 Vision 90B in the Magnetic-One Framework), reducing computational costs by 35%.
- Introduced a comprehensive web content extraction and summarization pipeline utilizing Playwright, integrating a RAG system that fetched user query outputs 50% faster.


C&S Electric Ltd.....
AI Intern June 2024 – July 2024, Noida

- Built and deployed a RAG application using AWS, reducing embedding time from 15 minutes to 30 seconds and query time from 2 minutes to 5 seconds.
- Executed a serverless architecture with Docker and AWS Lambda, streamlining asynchronous query handling and background processing.
- Led the incorporation of session IDs into the model’s architecture, enabling efficient real-time query data storage.

PROJECTS

Python Autocomplete System with VSCode Extension.....
Oct 2024  [Code](#),

- Developed an AI-powered autocomplete system leveraging **PyTorch** and **Transformer-XL** to achieve sub-100ms latency and improve prediction accuracy by 15% through optimized multi-threading and beam search.

Phishing URL Detection System.....
Aug 2024  [Code](#),

- Developed a phishing URL detection system capable of analyzing over 100k URLs per batch with real-time predictions (>1k/min), reducing processing time by approximately 40% compared to previous models.
- Integrated automated CI/CD pipelines using Docker, AWS, Airflow, and Terraform to streamline deployment.

POSITION OF RESPONSIBILITY

Google Developers Group OnCampus IIITN.....
AIML Core Team Member 2024 - Present,

Discipline Committee, Abhivyakti.....
Co-lead 2025,

SUMMARY

AI/ML enthusiast with hands-on experience in developing robust AI solutions. Skilled in designing multi-agent frameworks, deploying RAG applications, and streamlining system architectures. Experienced in solving complex AI/ML challenges through innovative modeling and optimization techniques.




SKILLS

| | |
|-------------------|--|
| Languages | Python, C++ |
| Libraries | Pytorch, Langchain |
| Frameworks | Flask, LangGraph, Autogen |
| Databases | MongoDB, Neo4j, Chromadb |
| DevOps | Docker, AWS |
| AI/ML | Generative AI, Deep Learning, NLP, Computer Vision |
| Additional | Multithreading, API Development, Debugging & Logging |

EDUCATION

| | |
|----------------------------------|----------------------|
| IIIT Nagpur | Nagpur, India |
| <i>B.Tech, CSE (AI & ML)</i> | <i>2022 - 2026</i> |

CERTIFICATIONS

| | |
|---|------|
| NVIDIA: Fundamentals of Accelerated Computing with CUDA Python | |
|  Certificate | 2025 |
| NVIDIA: Transformer-Based NLP Applications | |
|  Certificate | 2024 |
| NVIDIA: Fundamentals of Deep Learning | |
|  Certificate | 2024 |