Vishesh Chahar

Phone: (+91)-9501006533

Email: vishesh.chahar01@gmail.com
LinkedIn: www.linkedin.com/in/visheshchahar
GitHub: www.github.com/Vishesh-Chahar
LeetCode: www.leetcode.com/vishesh_chahar

BIO

Data Scientist specializing in **Generative LLM Agents**, **Machine Learning pipelines**, **Natural Language Processing**, and **LLM fine-tuning**. Proficient in **Python**, leveraging **CUDA-enabled PyTorch** for accelerated model training and optimization. Experienced in **ELT pipeline scaling** using **Kafka** and **PySpark**, and developing adaptive **RAG architectures** with **LangGraph** and **FAISS**. Skilled in building secure, scalable **REST APIs** with **FastAPI** and **Flask**, and in fine-tuning transformer models using **LoRA** and **PEFT**. Adept at deploying containerized ML systems with **Docker** and automating version control workflows with **Git**.

EDUCATION

Thapar Institute of Engineering and Technology, Patiala, Punjab
Bachelor of Engineering in Computer Engineering — Oct 2020 - Jun 2024

Bhavan Vidyalaya, Chandigarh
Senior Secondary School — Mar 2018 - Mar 2020

St. John's High School, Chandigarh
Secondary School — Mar 2006 - Mar 2018

EXPERIENCE

Ntigra AI Applications and Services LLC, Dubai, UAE — *Mar 2025 - Oct 2025*
ML Engineer

Clinical Temporal Relation Extractor — *[Python, Transformers, PEFT, CUDA, Pydantic]*

- Demonstrated a **40% increase in accuracy** in clinical entity coding using a novel temporal data extraction approach for clinical notes
- Improved timeline accuracy by **31% (F1: 0.67 → 0.88)** through hybrid **rule-based + transformer** logic using spaCy and regex-driven normalization.

Ntigra Medical Agent Backend — *[FastAPI, LangGraph, GLiNER, FAISS, HuggingFace, Vosk]*

- Architected a modular **agent-tooling framework** with **LangGraph Agentic State Workflows**, integrating **DuckDuckGo search, CSV parsing, and memory persistence nodes**, enabling dynamic multi-tool orchestration and error-tolerant workflow execution.
- Delivered **99.2% uptime** for **FastAPI backend** powering Ntigra's custom end-to-end **clinical AI agent**
- Reduced **latency by 46%** using asynchronous WebSocket streaming with **Vosk ASR**, and optimized conversational recall via **LangGraph + FAISS memory**.
- Improved entity precision by **34%** using hybrid extraction (**GLiNER + LLM memory**), automating clinical NLP workflows previously requiring manual annotation.

Isourse Technologies — *Jun 2024 - Dec 2024*
AI/ML RND Team

Gen-BI — *[Llama 3.1, PostgreSQL, Flask, Python]*

- **Led** a cross-functional RND team to develop a novel BI module with **Llama 3.1** reducing skill dependence and manual reporting by **20+ hours/month**
- Crafted and maintained secure, scalable APIs using **Flask** supporting over 1000 concurrent users, leading to reduced application latency

Data Warehousing — *[Phi-4, Apache Kafka, PySpark, Docker, K-Means]*

- Lowered data storage and processing costs by **30%** with a hybrid Data Warehousing solution integrated with **Phi-4, K-Means**
- Improved pipeline efficiency by **42% (74s to 43s)** and transformation speed by **34% (37s to 24s)** by using **Kafka**-based data streaming and optimizing **Spark Executors**.

Computer Vision — *[YOLO, Pytorch, CNN, LSTM]*

- Accomplished **83% accuracy on a custom CNN for handwritten OCR** on multi-digit numbers with a model size of 6.7 MB
- Attained **66.4% accuracy on object detection task** using **YOLO** for document digitization on self-annotated dataset

Wipro Limited — *Jan 2024 - Jun 2024*
Data Science Intern — *[BERTForQA, Llaama 3, Scikit-Learn, Pandas, BeautifulSoup]*

- Decreased data preprocessing delays by **13%** by implementing automation script for data extraction using **BeautifulSoup**
- Minimized **manual work hours by upto 30 hours/month** by automation of data extraction and processing using **Pandas**
- Increased system reliability by utilizing **AWS Sagemaker** to containerize pipelines
- Achieved **r2 score of 0.92** with **regularized in-house regression model** for prediction tasks

PROJECTS

Headliner — *Project Link* — *Python, CUDA, NLP, flan-t5*

- Fine-tuned **27.36%** of parameters of **flan-t5-large** with **LoRA** for headline generation from 236 articles with over 600 tokens each.
- Attained a **150% decrease** in model training time by implementing GPU computations using **CUDA**.

Diagnosis Pal — *Project Link* — *Python, ML, CategoricalNB*

- Strategized data preprocessing techniques for disease classification using **CategoricalNB** from symptom data
- Enhanced predictive precision, achieving a **84% MAP@K score**, and improved to **92% MAP@K** with hyperparameter adjustment.

SKILLS

- **Languages:** Python, R, C++, C#, SQL, Bash, Shell
- **AI/ML:** PyTorch, TensorFlow, Keras, CUDA, Scikit-learn, LoRA, PEFT
- **NLP:** LLaMA, Qwen, BERT, FLAN-T5, Hugging Face, NLTK, spaCy
- **CV:** YOLO, CNN, LSTM, OpenCV
- **Data Engineering:** Kafka, PySpark, PostgreSQL, FAISS, Pandas
- **CI/CD & APIs:** FastAPI, Flask, Git, Docker, Postman, LangGraph
- **Speech & Agents:** Vosk ASR, LangChain, GLiNER, Wav2Vec2