Jay Singh

Jays.iitkgp@gmail.com | +91-8651274328 | LinkedIn | Github

EDUCATION

IIT Kharagpur **2021-2025**

B.Tech Metallurgical and Materials Engineering (Minor in AI) *Kharagpur, West Bengal*

- **Relevant Coursework:** Probability and Statistics, Machine Learning Foundations, Applied Mathematics, LDA.

WORK EXPERIENCE

Innovaccer **June 2024 – Aug 2024**

Data Science Intern *Noida, UP*

- Amplified the natural language semantics of "Sara", a med-**AI model**, as part of NLP R&D, with accuracy **85%.**
- Harnessed **LangChain** for **COT, ZSL, few-shot RAG**; outperformed prior methods, cutting latency by **73%**.
- Developed a multi-chain asynchronous architecture for diverse use case segmentation using **GPT-4o** and **SLMs**.

Affine **April 2024 – June 2024**

Machine Learning Intern *Bangalore, Karnataka*

- Formulated a T5 220M model using FP16 checkpointing, reducing GPU usage by **55%**, efficient A100 training.
- Fine-tuned with LoRA, adding 2% trainable params, boosting BLEU/ROUGE-L by 1.7, reducing GPU usage.
- Launched an auto-scaling Kubeflow–MLflow–Triton pipeline, driving **10x** user growth, **30%** lower cloud usage.

Cambridge Judge Business School **May 2023 – Sep 2023**

Research Apprenticeship *Remote*

- Innovated an AI Startup Portal via **Groq API**, halving manual research time and generating **100+** business ideas.
- Engineered **NLP** parsers for extracting and transferring insights, generating SWOT analysis, boosting efficiency.
- Implemented scalable **RESTful** endpoints with CORS, handling over **200** requests per day, with **99.9%** uptime.

PROJECTS

Medium Article Search and Summarizing Agent **Jan 2024**

- The Medium summarizer generates concise summaries, with **5s query latency**, cutting research time by **70%.**
- Leveraged LangChain**, Anthropic, Hugging Face's** model **Flan T5 XL** and other LLMs for response parsing.
- Deployed over Render; the low-latency **Streamlit** UI provides a seamless user experience, and efficient caching.

Music Popularity Estimation **Jan 2024**

- Programmed a Music Popularity Estimation model with an MSE of **3.2**, R^2 score-**0.85**, over metadata features.
- Executed PCA on 2.2K-song metadata, reducing features by **35%** while preserving **90%** variance and R^2 of 0.85.
- Trained and optimized Random Forest, Gradient Boosting models, for boosting accuracy by **15%** over baseline.

CERTIFICATIONS & SKILLS

- **Certifications:** PyTorch for Deep learning, Career Essentials on Gen AI, Advanced Learning Algorithms
- **Technologies/Frameworks:** AWS(S3, Athena, Sagemaker), Snowflake, Flask, FastAPI, Django, Docker, SQL Kubernetes, MLflow, Kubeflow, RAG, ReAct, Git, GitHub, TensorRT, Plotly, Streamlit, Gradio, GPT, BERT
- **Skills:** Python, TensorFlow, PyTorch, Scikit-learn, Transformers, AutoML, NLP, MLOps, Federated Learning, Information Retrieval, Tf-IDF, Search Ranking Recommender Systems, Statistical Analysis, Analytical Thinking

COMPETITIONS & HACKATHONS

Amex Decision Science Track (Global Pre-Finalist, top 0.5% of 6,000+ teams) **June 2024 – July 2024**

- Ran ensemble models (XGBoost, LightGBM, CatBoost) on **100K+** T20 match records, achieving **92%** accuracy.
- Reduced prediction error by **15%** via **GridSearchCV** and **10-fold CV**, outperforming baseline methods by **20%.**

Impetus 5.0 Case Study (National Semi-Finalist) **Nov 2023 – Dec 2023**

- Crafted a data-driven business model, GTM strategy to drive investment and a **20%** increase in user acquisition.
- Performed K-means clustering on **5,000** customer survey responses, informing Finvest Pro's targeted marketing.