Krishanu Das Baksi

krishanu.dasbaksi@outlook.com
India

EDUCATION

2016
Indian Institute of Technology Bombay | Mumbai, MH
Master of Technology: Biomedical Engineering

2012
West Bengal University of Technology | Kolkata, WB
Bachelor of Technology: Biotechnology

CORE SKILLS

- Python
- Natural Language Processing
- Deep Learning
- Machine Learning
- Large Langauge Models
- Generative AI
- Vector Databases
- Transformers
- Computer Vision
- Knowledge Graphs

DETAILED SKILLS

- **Python Libraries and Tools**: PyTorch, Transformers (Huggingface), vLLM, Langchain, LlamaIndex, llama.cpp, Scikit Learn, Matplotlib, Pandas, NumPy, SciPy, SpaCy, NLTK, OpenCV, Flask, StreamLit, and many more..
- **Other Programming and Query Languages**: C++, JS, R, SQL
- **Vector Databases**: QDrant, Milvus, FAISS
- **Other Technologies**: Linux (& shell scripting), SQL, RDF, SPARQL, GraphDB, AWS, GPU Computing, Git, MLOps, Web Development.

PROFESSIONAL SUMMARY

A dynamic and results-driven Lead Data Scientist with a passion for leveraging cutting-edge technologies to drive data-driven decision-making. With a track record of delivering innovative solutions, I have the ability to rapidly adapt to new challenges and stay at the forefront of emerging trends in the ever-evolving field of data science. As both a hands-on coder and a contributing leader, I dive deep into technical details while actively contributing to solution decision-making, as well as providing guidance and mentorship to team members.

EXPERIENCE

05/2023 - Current
Lead Data Scientist
Deloitte (USI) AI CoE | Gurgaon, India

- Specialized in Natural Language Processing (NLP) and Generative AI as the primary fields of expertise.
- Developed NLP solutions using cutting-edge technologies including transformers, vector databases, large language models (LLMs), and generative AI.
- Developed RAG based Agent-like chatbots for conversing with users and answering complex questions.
- Led and managed teams of 2-3 data scientists.
- Collaborated with cross-functional teams to define project goals and objectives.
- Conducted research and stayed updated on the latest advancements in data science and AI to continuously enhance the team's capabilities.
- Provided mentorship and guidance to junior data scientists, fostering their professional growth.
- Published a peer reviewed research paper in highly reputed venues for NLP.

07/2016 - 04/2023
Researcher
TCS Research | Gurgaon, India

- Developed innovative AI-based solutions for addressing complex challenges in the field of life sciences.
- Focused on a wide range of problems spanning Natural Language Processing (NLP), computer vision, knowledge graphs, and other AI domains.
- Applied advanced AI, primarily NLP and Computer Vision models, for solving biomedical problems such as diagnosing celiac disease, creating structured knowledge from unstructured text etc.
- Contributed to the development and maintenance of knowledge graphs to enhance data-driven research in the domain of microbiology.
- Published multiple peer reviewed research paper in highly reputed venues.
- Led teams of up to 2 researchers to work on specific research projects, providing guidance and coordination.

- **Machine Learning**: Thorough knowledge of machine learning algorithms for regression, classification and clustering.
- **Deep Learning**: In-depth grasp over deep learning algorithms and techniques like Transformers, Convolutional Neural Networks, Feed Forward Neural Networks.
- **NLP**: Classical models (TFIDF, BOW) and neural models (LLM, transformers, LSTM, GRU, RNN etc.), language models (BERT et. al.), word representations, information retrieval (ElasticSearch) etc.
- **Generative AI and LLMs**: Expertise in fine tuning LLMs like Falcon, Llama etc. for domain specific applications using methods like LORA, QLORA etc.
- **Prompt Engineering and Generative AI Solution Architecting:** Advanced knowledge of prompt engineering along with ability to design and implement advanced LLM architectures including developing agents.
- **Solution Building and Deployment**: Can build solutions end to end using Rest APIs as well as basic front-end development using StreamLit. Additionally, have good grasp of Cloud technologies, especially AWS. Regularly use Docker and getting exposed to Kubernetes.
- **Knowledge Graphs and Semantic Web**: Experience of creating and utilizing knowledge graphs.
- **Computer Vision**: Knowledge of traditional and modern computer vision techniques.
- **Data Science**: Knowledge and understanding of probability, linear algebra, statistical methods, data analysis and visualization techniques.
- **Skills currently Acquiring**: Kubernetes, Azure, LLMOps.

LANGUAGES

English, Bengali, Hindi: First Language

Spanish: A2

Elementary

Italian: A1

Beginner

SELECTED PROJECTS

Development of Scalable Agent based ChatBot using LLMs for SDLC Management (Area: NLP)

- Led a team in the development of a scalable chatbot using the Retriever Augmented Generation principle to answer SDLC related questions using data collected and maintained data by client team.
- Explored and implemented chatbots with agent-like capabilities using dynamic routing and function calling.
- Designed a microservices architecture for solving this problem at scale. Explored and utilized the fastest libraries for serving LLMs as well as vector databases for storing data.
- Packaged entire solution into multiple scalable APIs and actively participated in the deployment of the above solution on AWS EC2 Instance using Docker, and on Kubernetes cluster.

Identification of Requirements from Cloud Security Documents (Area: NLP)

- Developed a Retriever-Reranker architecture for mapping cloud security requirements from documents to a predefined set of integrated requirements.
- The developed methodology saves up to 80% of time needed for the task

Relation Extraction using Transformer (BERT) and Transfer Learning (Area: NLP)

- Developed transfer learning based method (using BERT) for named entity recognition and information extraction from scientific text.
- Led a team for thecreation of data annotation pipelines to help in annotation and review of data.

Question Answering using Knowledge Graphs (Area: NLP, KG)

- Developing an unsupervised NLP method for automatically answering biomedical natural language questions using the Wikidata knowledge graph.
- Developed unsupervised named entity recognition, entity linking and relation linking methods using multiple heuristics to convert natural language questions to logical queries.
- Method outperforms other SOTA methods on the task.

Automatic identification of Celiac Disease from Medical Images (Area: CV)

- Developed a method for identifying celiac disease from biopsy images using existing CNNs architectures using a mixture of novel CNN architectures.
- Reduced time for disease diagnosis by up to 90%.

CERTIFICATIONS

- **Deep Learning**, a 5-course specialization by **deeplearning.ai Coursera**
- **Data Science Applied Data Science with Python**, a 5 course specialization by **University of Michigan Coursera**
- **Accelerated Computing using CUDA** (C++ and Python), **Accelerated Data Science using CUDA - Rapids** by **NVIDIA**

RECENT PUBLICATIONS

1. Exploration of Open Large Language Models for eDiscovery, Pai et. al.; Proceedings of the Natural Legal Language Processing Workshop, EMNLP 2023;
Link: https://aclanthology.org/2023.nllp-1.17.pdf
2. BactInt: A domain driven transfer learning approach for extracting inter-bacterial associations from biomedical text, Baksi et. al.; Computational Biology and Chemistry;
Link: https://doi.org/10.1016/j.compbiolchem.2023.108012