

AI/ML Architect specializing in **NLP, LLMs**, and scalable deep learning systems. Expertise in **Generative AI, model fine-tuning**, and deploying large-scale production-ready AI solutions

## PROJECTS

---

### ❖ Salary Extraction & Normalization Engine LLM + NER Pipeline for Salary Structuring

- Finetuned a **RoBERTa**-based NER system to identify salary entities from 400M+ job descriptions at scale.
- Fine-tuned **Gemma-270M**, a decoder model to convert extracted salaries into normalized JSON format with range, currency symbol and ISO code.

### ❖ CURIE – Intelligent Assistant for Actionable Insights - (*A **flagship model** from DRAUP*)

- Understands user queries & intent using custom models, performing deep searches across multiple **RAG applications (30M+ embeddings)** with Milvus and several **Text-to-SQL systems** to extract insights (**2M+ table entries**).
- **Combines RAG and Text-to-SQL outputs** using **Chain-of-Thought** reasoning with custom models and LLMs for context-aware insights. Optimizes real-time retrieval and response generation for accurate, intelligent assistance.

### ❖ LLM-Powered News Engine: Built a scalable news processing system similar to Google News by developing in-house transformer-based language models

- Continual pre-trained a **RoBERTa**-based transformer language model on (10M+ news) 550M tokens, reducing loss by 60% & improving various downstream NLP tasks like **classification, NER, semantic embedding, recommendations**, etc.
- Built **an in-house scalable clustering algorithm** to generate news trend recommendations at scale.

### ❖ **Profile Recommendation Agent:**

- Built a Reflection-pattern AI agent using LLM-driven filter generation with critique/validation loops to parse job descriptions and retrieve high-relevance matches from 400M+ Elasticsearch profiles.

### ❖ **Automated Job Responsibility Summarisation System:** Led the development of an automated system to extract and summarise job responsibilities for professionals.

- Trained **Seq2Seq BART model** for large-scale responsibility **summarisation**, removing the need for manual work.
- Created a custom embedding model for Job Description-to-Profession matching, improving work experience interpretation for a large user base.

### ❖ **Information Retrieval (IR) System for Domain-Specific URLs**

- Designed and implemented a machine learning system utilizing **XGBoost and Random Forest** algorithms
- Engineered multiple Machine Learning tools for **effective annotation** of training data, including the creation of an [HTML Tag Annotator](#)

## EMPLOYMENT

---

DRAUP Inc. (Oct 2018- present), **Data Science Architect**

- ❖ Lead a **12-member Data Science team**, driving the adoption of ML and deep learning technologies for real-world applications.
- ❖ Collaborate with **product, big data, and research teams** to develop AI-driven solutions for business and client needs.
- ❖ **Designed, deployed, and optimized** multiple ML/DL/LLM models in production on AWS ECS, leveraging GPUs and Apache Spark
- ❖ Built and scaled Retrieval-Augmented Generation (**RAG**) applications and Text-to-SQL systems, integrating **vector databases**, trained embedding models, and LLMs to enhance NLP **search, retrieval, and query** generation.

CALIENT Technologies Inc.,(Oct 2017 - Sep 2018), *Machine Learning Engineer (Software)*

- ❖ Built Machine Learning models to categorize network device error logs into multiple categories (**word2vec + logistic regression**) and implemented an alarm system using **Java** for quick response.

COGOPORT, (Feb 2017 - Sep 2017), *Software Developer*

- ❖ Designed and Developed Bid Request System including **data analytics pipeline** and API development using **Ruby on Rails**.

TECHNICHE e-commerce solutions, (Apr 2015 - Feb 2017), *Full Stack Developer*

- ❖ Proficiently designed and developed multiple **e-commerce websites** using **Node JS (MEAN stack)**, encompassing essential features such as user authentication, authorization, shopping cart functionality, and comprehensive order management

Hewlett-Packard(HP), (Aug 2013 - Dec 2014), *Software Engineer*

- ❖ Created **Java** applications for internal use, utilizing extensive SQL expertise for banking project

## Skills, Languages & Technologies

---

- PyTorch, HuggingFace 🤗, Milvus, W&B etc.
- ML & NLP: Neural Networks, LSTMs, Transformers, LLMs, Quantization, Embedding Models etc.

## EDUCATION

---

Bachelor of Engineering (B.E.) in Computer Science (2009 - 2013) - VTU