AAINA BAJAJ

(Senior Data Scientist | PAHM®)

90230 - 99769 | aainabajaj220@gmail.com

https://www.linkedin.com/in/aainabajaj | https://medium.com/@aainabajaj39

PROFESSIONAL SUMMARY

I have 6 years of experience in data science with advanced technical proficiency, a deeper understanding of the healthcare insurance domain, managerial and strategic thinking skills. I am part of the core Innovation Council team where we have designed and developed various NLP-based systems, including Document Summarization, Logical QA system, Information extraction system, semantic search, and chatbots. I have finetuned and evaluated open-source LLMs with our custom dataset. Techniques include LORA, QLORA, and quantization techniques. I am experienced in leading cross-functional teams of data scientists and engineers and achieving quality work. Excellent communicator and collaborator, able to convey complex concepts to diverse stakeholders.

TECHNICAL SKILLS

- **Generative AI / Large language models**:
  - ✓ Varied experience in leveraging Langchain, GPT models, and open-source LLMs for different use cases.
- **Finetune LLM techniques:**
  - ✓ Finetuned and evaluated open-source LLMs with our custom dataset. Techniques include LORA, QLORA.
  - ✓ DPO on text summarization.
- **Model quantization techniques:** GPTQ (post-training quantization), bitsandbytes (dynamic quantization)
- **In-depth knowledge of advanced architectures:** Includes Attention, Transformer, BERT, and GPT models.
- **Strong hands-on various Embeddings**: InferSent, Universal Sentence, Glove, Sense2vec, Fuzzy, Tiktoken, and Huggingface-based model embeddings.
- **Strong hands-on various architectures**: including Siamese networks.
- **All Machine Learning and Deep learning models**- For regression, classification, EDA, outlier detection & more.
- **Intuitive in text pre-processing and normalization techniques**: tokenization, POS tagging, parsing, Dependency parser, Context-free grammar, and how they work at a low level.
- **Image preprocessing**: using OpenCV.
- **Strong with open-source NLP toolkits** such as Stanford CoreNLP, and OpenNLP.
- **Data collection and version handling**: Can handle PDF, XML, Images, Web, Emails; with versioning.
- **Data Annotation tool**: Prodigy.
- **Deployment/User Interface tools** - Docker, Flask, Gradio, Streamlit.
- **Databases and Languages**: SQL; Skilled in Python, and Java.
- **Secondary skills**: RDF, Pyke rule engine, Semantic Web, Knowledge Graph, Spark, Hadoop, Hive, Graph database-SPARQL, and Kubernetes.

WORK EXPERIENCE

Senior Data Scientist III, 03/2023 to Current

Carelon Global Solutions - Hyderabad, Telangana.

- Implemented **Information extraction system**, **text summarization,** and **chatbots** using LLMs:
  - Handled certain logical questions like finding family relationships from the medical documents which the normal language models are not able to handle. Hence, resolved issue using LLMs
  - Finetuned and evaluated open-source LLMs with our custom dataset, which outperformed OpenAI's GPT 3.5 models.
  - Used LORA - parameter efficient finetuning so that the model can be fine-tuned on lower resources. Further quantized model using GPTQ to reduce latency time.
  - Applied retrieval and reranking augmented generation (RAG) to extract relevant chunks from documents.

- Varied experience in leveraging LLMs has led to mastery in prompt engineering.

- Led a cross-functional team of data scientists and engineers and ensured the project's success.
- Led innovative initiatives and mentored junior data scientists, providing guidance, and fostering their professional growth**.**
- Present findings and recommendations to senior management, translating complex technical concepts into actionable insights.
- Currently, Finetuning opensource LLM using the DPO method on **text summarization.**

Data Scientist II, 05/2021 to 02/2023
Carelon Global Solutions - Hyderabad, Telangana.

- Implemented **Automatic Document to Medical Necessity Scenarios Conversion System, Custom NER model** and **Automatic Decision System:**
  - Automated critical manual process of getting medically necessary scenarios from policy medical guidelines.
  - Single-handedly designed and developed complete solution.
  - It reduced the team's efforts by ~50% and improved production by 4.5 times.
  - Dependency parser combined with NER technique and custom logic, was the core of system.
- Deployed project in production in collaboration with engineers.

Data Scientist, 07/2019 to 04/2021
Vertogic LLP (renamed as 'Sapvix India LLP') - Hyderabad, Telangana.

- Engineered end-to-end information extraction pipeline, using Spacy and graph database. It includes the collection of data from different formats and the creation of a custom PYKE rule engine and query using SPARQL queries.
- Developed BERT-based QA system and deployed through Flask interface.
- Led the team and handled the project's overall progress.

Mainframe Developer (Insurance Domain) - Software Engineer, 07/2017 to 08/2018
NIIT Technologies - Greater Noida, UP.

- Identified production issues impacting code modules. Ensured accurate issue tracking and reporting.

EDUCATION / CERTIFICATIONS

Professional, Academy for Health Care Management (PAHM®): Earned designation with a score of 91%. (Oct 2023)
AHIP- America's Health Insurance Plans: Online

PGP In Big Data Analytics and Optimization (Sep 2018-Mar 2019)
International School of Engineering (INSOFE)- Hyderabad, Telangana.

B.TECH/CSE: Graduated with an average of 83%,(2013-2017)
Punjab Technical University: Punjab

ACHIEVEMENTS

- Got '**Go Above awards**' for outstanding performance - 5 times in the last 2 years.
- Won scholarship based on performance at INSOFE, Hyderabad.
- Won another scholarship in Analytics Hackathon 2019 held under INSOFE.
- Cleared GATE 2017.
- One of the finalists in Smart India Hackathon 2017 held under Govt. of India.