AMBUJE GUPTA
Email: ambujegupta99@gmail.com || Contact No: 9999715392

Patent - EEG
GitHub – github.com/ambuje
LinkedIn - www.linkedin.com/in/ambuje/
Medium - ambuje.medium.com/
Google Scholar - scholar.google.com/citations?user=ZU2fXr4AAAAJ&hl=en

EDUCATION

Bennett University, Greater Noida 2017 - 2021
Bachelor of Technology, Computer Science Engineering *CGPA (9.18/10)*

TECHNICAL SKILLS	TRAINING AND CERTIFICATIONS
• **Programming** : Python **Database** : MySQL • **Software Package** : Microsoft Office, G-Suite, Postman • **Tools/Technologies** : Machine Learning, Deep Learning, NLP, TensorFlow, Keras, Pandas, NumPy, Transformers, Bert, NLU, LLM, LLAMA, LORA, Mistral, GPT, LSTM, Classification, Generative AI, Sagemaker, Distributed Training, Accelerate, Deepspeed	• Neural Network and Deep Learning (Andrew NG) (Coursera) (2020) • Sequence Models for Time Series and Natural Language Processing (Coursera) (2020) • Continuous Delivery & DevOps (Coursera) (2020) • Fundamentals of Parallelism on Intel Architecture (Coursera) (2020) • Attended a Three-day workshop on Big Data and Machine Learning (Bennett University) (2017)

Work Experience

Uniphore, Bengaluru

• AI Scientist (Data Science) March 2022 – Present

1. KGRag (Knowledge Graph Retrieval-Augmented Generation Framework)
   a. Designed and implemented an advanced RAG system leveraging knowledge graphs for enhanced information retrieval in large language model applications. The framework utilizes graph-based document representation to provide more contextually relevant information to LLMs, improving answer accuracy by **5%**
   b. Architected a modular framework with separate **design-time and runtime components** for knowledge graph creation and query processing
   c. Engineered **streaming data support** for unlimited incremental processing of new documents with efficient path reuse, enabling continuous knowledge base expansion without reprocessing existing content
   d. Developed a multi-stage retrieval pipeline with query classification, vector similarity search, and cross-encoder reranking
   e. Created flexible database abstractions supporting both **local/S3-based LanceDB and cloud-based AstraDB storage**
   f. Optimized **asynchronous** processing with concurrency control for embedding generation and LLM interactions
   g. Finally created a API using **FASTAPI**
   h. **Technologies**: Python, LLMs (Llama-3.3-70B), Vector Databases (LanceDB, AstraDB), AWS S3, Knowledge Graphs, Embedding Models, Retrieval-Augmented Generation, Cross-Encoders
2. Working on Agentic AI
   a. Developing an agentic AI service implementing **Pre-Act methodology**, which enhances agent performance through **multi-step execution planning with detailed reasoning for each action**
   b. Utilized LLMs as the core framework with Langgraph for orchestration, enabling incremental plan refinement after each step execution
   c. Implemented feature engineering on function calling datasets, **achieving 94% accuracy** on Glaive dataset, outperforming GPT-4 through strategic fine-tuning of smaller models (8B & 70B)

d. **Designed a two-level evaluation framework**—turn-level (e.g., Action Recall) and end-to-end—augmented with business KPIs (goal completion rate, progress rate). The Goal competition and progress rate metric helped to evaluate the system on business use cases.
e. Paper - Pre-Act: Multi-Step Planning and Reasoning Improves Acting in LLM Agents

3. Developed Semantic Match Configuration
   a. Designed and implemented a system for analyzing conversations based on key phrases
   b. Used SPARSE vector search (**SPLADE**) to have the initial k turns related to the key-phrase and then used a **Re-Ranker** (BGE) to have the final output
   c. SPLADE helped to achieve high recall, but poor precision and re-ranker helped us to achieve high precision, hence the overall system achieved a good **85% accuracy** (high recall, high precision)
4. Worked on El-Lifecycle (Low Code LLM Training and Inference Module)
   a. Developed a low-code tool designed to efficiently and effectively train Large Language Models (LLMs) on Amazon SageMaker using Python.
   b. Simplified user experience by requiring only data and hyperparameters, with the tool handling the rest of the process.
   c. Implemented best 'x' model selection mechanism based on ROUGE scores instead of evaluation loss, followed by end-to-end testing with the specified evaluation method on all the 'x' models. Final output is the best model
   d. In the tool, we have implied distribution training for faster training and inference
      a. Training – Integrated accelerate and deepspeed with ZeRO stage 2, about **8x faster** than normal distributed training
      b. Inference – End to end inference is integrated with VLLM, it makes inference **24x faster** than the traditional or batched inference.
   e. The tool not only helps in training and evaluation but also helps in managing the model lifecycle
5. Working on Retrieval Enhancement through Domain-Adaptive Fine-Tuning via Model Fusion of Embedding Models
   a. Developed a model fusion approach for fine-tuning domain specific data to have better contextual embeddings.
   b. The system aims to increase overall accuracy of Retrieval in a QA system.
   c. The developed system outperforms the current system (fine-tuning BGE embeddings) and also reduces the problem of catastrophic forgetting.
   d. Have been working with the **BGE (SOTA) embedding for experimentation**.
   e. The retrieval performance are measured using 4 metrics :- MAP, NDCG, MRR, Recall.
   f. Paper - **REFINE on Scarce Data: Retrieval Enhancement through Fine-Tuning via Model Fusion of Embedding Models**
   g.
6. Working on Zero/Few-Shot Entity Classification
   a. Using **LORA**, and **PEFT** techniques to instruct **finetune 7B (LLAMA, Mistral) models**.
   b. Aim **to instruct finetune** the model with larger dataset which can capture entities in broader aspect
7. Worked on Zero/Few-Shot Intent Classification
   a. Continual Pre-training Language Models on Domain Specific data using infilling token technique.
      a. Ssupervised and unsupervised both ways are used to evaluate pre training.
      b. In Unsupervised, made clusters and used hungry alignment to align the clusters
      c. Final 3 metrics are used to evaluate the clusters i.e. ACC, ARI, NMI and saw a good jump of 80 percent.
      d. Also checked if there is **catastrophic forgetting.**
   b. Different techniques like **NLI, SetFit** to train the classification model.
   c. Using **Llama** and other LLMs to predict intent without training.
   d. Finetuning LLMs using LORA, on a downstream task.

8. Experienced in developing a Large Language Model with a focus on contextual understanding.
   a. Utilizing Decoder models, such as UL2, combined with prompt engineering to effectively capture contextual entities.
   b. Successfully tackling the challenge of working with limited available data.
9. Data Augmentation – Working on LLMs (10B+ param) to generate data. Using LLAMA 2, Flan UL2 for this task.
10. Proficient in leveraging question answering techniques, employing two classifiers, to effectively capture contextual entities.
11. Well-versed in implementing online learning methodologies for classification tasks.
12. Designing Generative AI life cycle for these projects.

• Associate AI Scientist (Data Science)

1. Worked on Slot filling.
   a. This project combined clubbing system entities (cardinal, person, loc, org, money etc) with the context

b. For context we infused system entity information at the classifier level, i.e. playing with how to take embeddings for our use case.
c. For data augmentation I have been researching Large Language models like Flan T5, Ul2 etc.

2. Developed a rule (Phone Number, Insurance number and Date of Birth) and **spacy** (Person name, Email ID) based Name entity recognition (NER) model for Japanese clients. For NEC (Japanese client), I was involved in end-to-end development and testing (Postman) which eventually helped the company in onboarding the client.
3. I have also developed **"Intent Based Classification"** for an Australia Bank which would help them to classify intent of a customer for their products. Moreover, we are also providing the words in a sentence that are responsible for that classification.
    a. **Transformer (Distil Bert)** has been used to for classification.
    b. Spans (words) in a sentence have been picked up by using the **attention score** from the last attention layer of the model.
4. Lately, I have also developed a **transformer-based NER model** for the Arabic Client. **Camel Bert** model has been fine tuned for Person and Organization Entity.
    a. With my approach, I brought about **90.4762%** change in Person and **666.667%** change in ORG entity as compared to their previous production model. (%Change in the F1 score)
5. I was also involved in rule based **Inverse Text Normalization (ITN)** for Malayalam Language. I was able to develop a system with more than 90% accuracy.

ZS Associates, Gurugram December 2020- March 2022

• Business Technology Solution Associate

1. Working in the R&D of BMS. Responsible for automating process for more efficiency.

RESEARCH EXPERIENCE, PUBLICATION AND PATENT

Pre-Act: Multi-Step Planning and Reasoning Improves Acting in LLM Agents May 2025

Introduced Pre-Act, a novel extension to the ReAct paradigm that generates explicit multi-step execution plans alongside detailed reasoning, iteratively refining each step based on prior actions and tool outputs

REFINE on Scarce Data: Retrieval Enhancement through Fine-Tuning via Model Fusion of Embedding Models October 2024

Proposed REFINE, a two-stage approach that (1) synthesizes in-domain training pairs from available documents and (2) fuses multiple pre-trained embedding models into a single fine-tuned retriever—boosting low-resource retrieval without sacrificing out-of-domain generalization

Bennett University, Greater Noida

Decryption of Brainwaves for Thought To text

Researcher

- Worked on fabricating the motor for especially abled people; fabricated a prototype of a wheelchair, its movement could be controlled with the subject's thought; the EEG device work by the person would capture EEG signals.
- Received ***seed funding*** from the university upon showing the first prototype.
- As our first aim was to solve mobility issue it was vital that the EEG device the subjects put on should be handy i.e., should be small. I have worked on 4 EEG headsets in which 3 headsets are small and can be easily put on the head.
    - Neurosky Mindwave Mobile 2 (1 Channel)
    - Muse (4 Channel)
    - Emotive Epoc + (14 Channel)
    - 32 Channel EEG device (Bulky)
- Channels here could be understood as the value of n nodes given by the subject's brain (n represent number of channel) • The technologies used is Deep Learning. Specifically, LSTM, CNN+LSTM is used and all of these are coded in Python.
- Filed a ***patent*** on November 17, 2020; in process of publishing the findings in a refereed journal

NLP-Titan at SemEval-2023 Task 6: Identification of Rhetorical Roles using Sequential Sentence Classification March 2023

- The paper presented a novel way to identify Rhetorical Roles in a legal document.
- This task is like sequential sentence classification. Used BERT HSLN (Hierarchical Sequential Labeling Network) algorithm to complete this task. On BERT HSLN we applied SetFit to have better accuracy. SetFit is applied to classes which have less data. The task is coded using python and pytorch is used to implement the algorithm. The architecture consists of transformer (BERT) for word embedding and Bi-LSTM that help us to have sentence level relation in a long document.

BennettNLP at SemEval-2021 Task 5:
Toxic Spans Detection using Stacked
Embedding Powered Toxic Entity Recognizer August 2021

- The paper presented a novel way to class toxic spans in a sentence.
- Used concepts of deep learning Stacked Embedding and Linked-list based pre-processing to accomplish the task. Got an opportunity to write a paper in ACL**-IJCNLP**, which comes under top 5 conference of NLP.
- Various NLP techniques as well as state of the art models are used in this paper. Specifically, Flair and Bert embeddings are used to generate the required result. All of these things are coded in python. Data manipulation work is achieved by using pandas and NumPy library of python.

BennettNLP at SemEval-2020 Task 8:
Multimodal sentiment classification
Using Hybrid Hierarchical Classifier December 2020

- The paper presented a novel way to class memes into19 categories
- Used a machine learning based hybrid classifier to accomplish the task. Got an opportunity to write a paper in **Coling**, which comes under top 5 conference of NLP.
- This problem is known as multi dimension data classification i.e., predicting various classes for a single piece of text.
- Concepts of Natural Level Processing (NLP) has been used to accomplish this task. Different pre-processing techniques include:-
  - Handling Imbalance Data
    - Data Manipulation
    - Removing unwanted Characters
    - Stemming
    - Lemmatization