Ambuj Krishna Agrawal

(412) 918-0594 | ambujagrawal741@gmail.com | LinkedIn | Website

EDUCATION

Carnegie Mellon University December 2025

Master of Science in Natural Language Processing (Artificial Intelligence), School of Computer Science *Pittsburgh, PA*

- **GPA: 4.0/4**
- **Coursework:** Advanced Natural Language Processing, Intro to ML, Multimodal Machine Learning.

Indian Institute of Information Technology, Allahabad July 2021

*Bachelor of Technology in Information Technology - (**Honors**: 8.82/10)* *Prayagraj, India*

- **Coursework**: Artificial Intelligence, Image And Video Processing, Linear Algebra, Data Structures and Algorithms, Distributed Systems, Database Management Systems, Operating Systems.

WORK EXPERIENCE

Carnegie Mellon University Pittsburgh PA

Machine Learning Research Assistant September 2024 – Present

- Working as a Research Assistant under Dr. Fernando Diaz, leveraging granular and implicit human preference feedback through infilling and edits to enhance personalization and reduce cognitive load in comparing LLM-based system outputs.

CRED Bangalore, India

Senior Software Development Engineer February 2023 – July 2024

- Built core components in the bill payments platform powering services like credit cards, electricity, rent, and gift vouchers. Built zero-to-one products that added revenue stream worth **$30 Million** a month and handled spiky traffic of more than 100 QPS.
- Solved a complex native memory leak in JVM due to memory fragmentation, cutting AWS cost by **25%** for multiple services.
- Developed a pipeline to deploy product config at CRED's hackathon saving **$150k** annually; team placed third in 70+ submissions.
- Intern: Made microservices from scratch for a new line of products around *Buy Now, Pay Later (BNPL)* collaborating across verticals.

LinkedIn Bangalore, India

Software Development Engineer July 2021 – February 2023

- Upgraded authentication cookie signatures to ECDSA-256 after testing on CPU, memory usage, and security benchmarks.
- Fixed major vulnerabilities like **replay** attacks during Sign in with Google flow by caching server-side nonce.
- Integrated sign-in with Facebook to move away from passwords with a **1%** WAU increase and collaborated across verticals.
- Optimized *last seen at* timestamp processing of users reducing resources by **50%** by using Samza streams and efficient batching.
- Intern: Re-architected the way of storing phone numbers. Also made a clustering algorithm to generate the required number of regexes for phone number validation to replace Google's "libphonenumber".

RESEARCH PROJECTS

Routers In LLMs October 2024 – December 2024

Carnegie Mellon University *Pittsburgh PA*

- Developed a novel method to obtain synthetic preference data between open-source models with scarce existing preference data. Built routers by fine-tuning Llama models and matrix factorization using the new data to enhance routing, tested on GSM8K.

Multimodal Web Agents September 2024 – December 2024

Carnegie Mellon University *Pittsburgh PA*

- Developed an ensemble of verifier agents using a tree search algorithm to address web agents' inefficiencies by pruning unnecessary paths, improving resource efficiency and accuracy. Benchmarked on the VisualWebArena dataset, enhancing the best-performing paper.

Semantic SLAM [Code] January 2020 – July 2020

Indian Institute of Information Technology, Allahabad *Prayagraj India*

- Improved localization accuracy of **V-SLAM** by injecting semantic information of detected corner points. **Published** as the first author in the Mediterranean Conference on Control and Automation (**MED**), Athens, Greece in 2022.

SKILLS

Programming Languages: Python; Java, C++, SQL, NOSQL, Go

Frameworks And Tools: PyTorch, Keras, Numpy, Pandas, Langchain, Flask, scikit-learn, NLTK, Kafka, AWS, Spacy

Areas of Expertise: Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), DS, DBMS

PROJECTS

RAG-Chatbot October 2024 – October 2024

- In a team of three, developed a RAG pipeline using concepts like multi-query, cross encoders, lost in the middle, BM25, and vector embedding with Langchain and scraped over 5000 documents using selenium to answer the latest questions about Pittsburgh.

liveAssist May 2024 – May 2024

- Built an assistant chatbot designed to interact with live sports videos, using Generative AI APIs provided by Gemini. Enhanced its performance, by incorporating Retrieval-Augmented Generation (**RAG**) to give targeted context to LLM to answer user queries.