

Curriculum Vitae/Resume

Ankit Kumar Pal

Principal Research Engineer

CONTACT INFORMATION

🌐 aadiya.github.io

✉ *E-mail:* aadiya@gmail.com

🎓 scholar 🐙 github 🌐 linkedin

PERSONAL INFORMATION

Date of Birth

June 2, 1996

Place of Birth

Sitapur, Uttar Pradesh, India

RESEARCH INTERESTS

Representation Learning on Graphs & NLP, Generative Large Language Models (LLMs), and their applications in Healthcare data, Federated learning, ASR & Audio Analysis

🎓 EDUCATION

Babu Banarasi Das University, Lucknow, India

May 2017

Bachelor of Technology, Computer Science Engineering

- **Thesis:** Generative Modeling of Music Sequences with LSTM-based RNN Architecture

Anandi Devi S.V.M, Sitapur, India (ADSVM), Sitapur, India

April 2013

12th - Board of High School and Intermediate Education U.P

- **Major:** Physics, Chemistry and Mathematics

🏢 EXPERIENCE

Saama Technologies, Chennai, India

May 2018 - Present

Senior ML Research Engineer

Objective: *Develop Deep Learning/NLP methods and pipelines for clinical data, Lead research projects, and published findings in top ML conferences*

- **Adverse Event Prediction, 2018:** *FDA Adverse Event Reporting System (FAERS)* Developed an RNN-LSTM model with Context-Aware Attention to extract pharmacological semantics from clinical notes, achieving 98% F1 score. Optimized character and word embeddings to enrich contextual representation. Enabled automated adverse event detection across 1M records.
- **Trial Plan Optimizer (TPO), 2018:** Designed a machine learning model using clinical trial data from a top-tier biopharmaceutical company to predict site enrollment. Implemented a Python and Scala AutoML framework with TransmogriAI. Utilized categorical embeddings and tree-based algorithms, including XGBoost, LightGBM, and Random Forest, which resulted in a 30% increase in the accuracy of site enrollment predictions, significantly improving trial planning and resource allocation.
- **Unsupervised Medical Monitoring, 2019:** Conducted analysis of clinical trial data across SDTM domains to identify patient outliers, ensuring early detection of anomalies that could impact trial results. Leveraged historical patient data and unsupervised models, such as Autoencoders, Clustering (e.g., K-Means, DBSCAN), Isolation Forest, and One-Class SVM, to optimize outlier detection. Implemented a human-in-the-loop process for feedback-driven model retraining. This iterative feedback process resulted in a 30% improvement in model accuracy, reduced false positives by 20%, and enhanced the reliability of clinical trial outcomes.
- **DeepMap ML Framework (SDTM Automap), 2020:** Developed a machine learning system to automatically generate CDISC SDTM mappings. The system incorporated Generative Adversarial Networks (GANs), Bidirectional LSTM with PubMed and BERT embeddings, and a 3-layer ELMo architecture for multi-task learning across clinical domains. It achieved an average

accuracy of 95% in mapping source raw data to SDTM standards, significantly reducing the time required to automate the mapping process.

- **Pharma Graph, 2021:** *Predictive Modeling of Drug Interactions using Graph Convolutional Networks* Built a NER model to extract pharmacological relationships from clinical text. Developed a Graph Convolutional Neural Network with attention mechanisms to model drug interactions from the clinical documents. The model showed a 20% improvement in miF1 over HSVM and 11%-19% accuracy improvement over hierarchical models like HEAGCRCNN, HAN, HiLAP, and HTrans, and a 16% improvement over Bi-BloSAN.
- **Large Language Models for Healthcare Domain, 2022-** Developed an internal Clinical LLM tasked with processing a variety of health-related data, which improved clinical insights extraction from raw protocol documents and PDFs. Utilized Retrieval-Augmented Generation to enhance information retrieval accuracy, reducing manual review time. Created a Python library for prompt versioning and structured outputs, which standardized and accelerated document generation processes. Generated protocol documents from minimal inputs, streamlining the documentation process by 50%. Conducted research to mitigate LLM hallucinations in the medical domain, resulting in a 10% decrease in erroneous outputs.
- **OpenBioLLM, 2023-2024:** Developed OpenBioLLM-70B and 8B, the most advanced open medical-domain Language Learning Models (LLMs) available. Achieved 86.06% accuracy on 9 diverse medical benchmarks, surpassing GPT-4 by 4% and other leading models. This model became the first healthcare model to trend on Hugging Face and the first Indian model to trend alongside Intel, Google, and Microsoft. Within five days of release, it gained significant traction and has been downloaded over 60,000 times in two months. Independently managed the entire project lifecycle, from data collection and cleaning to model training, alignment, and evaluation, demonstrating exceptional performance and impact in the AI and healthcare communities.

PROFESSIONAL EXPERIENCE

Prescience Decision Solutions, Bengaluru, India
Deep Learning Engineer

Feb 2018 - May 2018

Objective: *Building a Multidimensional Deep Learning Model to Predict the Bitcoin Price*

- Worked on transfer learning, attention methods, and custom POS-Tag embeddings.
- Developed a custom Twitter API to extract Bitcoin-related tweets and performed LSTM-based sentiment analysis. Incorporating sentiment scores as a feature layer in the prediction model increased prediction accuracy by 20%.
- Deployed the code and APIs, and built a Chat UI to facilitate interaction with the model, enhancing user engagement and ease of use.

Fliptango Global Solutions, Kerala, India
Machine Learning Intern

Dec 2017 – Feb 2018

Objective: *Design and implement an ML-driven e-commerce chatbot to optimize user interactions and enhance product recommendations*

- Used TensorFlow to leverage transfer learning and optimize models for specific tasks.
- Integrated Commonsense Embeddings from ConceptNet Numberbatch, enhancing the chatbot's understanding of language and increasing response accuracy by 25%.
- Followed the BiLSTM-CNN-CRF paper closely to build a named entity recognition (NER) model in TensorFlow, achieving 95% accuracy in identifying key entities from user chats, significantly improving the quality of product recommendations.

SELECTED PUBLICATIONS

Ankit Pal, Muru Selvakumar, Malaikannan Sankarasubbu. *Multi-label Text Classification using Attention-based Graph Neural Network. In Proc. ICAART, '20.*

Ankit Pal, Malaikannan Sankarasubbu. *Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing*. In **ACM '21**.

Ankit Pal. *CLIFT: Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain*. Poster in Robustness in Sequence Modeling Workshop **NeurIPS, '22**.


Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu. *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. In Proc. **PMLR '22**.

Madhura Josh*, **Ankit Pal***, and Malaikannan Sankarasubbu. *Federated learning for healthcare domain - pipeline, applications and challenges*. In **ACM '22**.

Ankit Pal. *DeepParliament: A Legal domain Benchmark & Dataset for Parliament Bills Prediction*. In Proc. **EMNLP '22**.

Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu. *Med-HALT: Medical Domain Hallucination Test for Large Language Models*. In Proc. **EMNLP Conll '23**.

Ankit Pal, Malaikannan Sankarasubbu. *Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations* In Proc. **NAACL Clinical NLP, '24**.

 IN PREPARATION **Ankit Pal**, and Pasquale Minervini and Andreas Geert Motzfeldt and Beatrice Alex. Open Medical-LLM Leaderboard. **Huggingface, '23**.

Ankit Pal, Malaikannan Sankarasubbu. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences **Under work, '24**.

+ SERVICE

Reviewed Papers for NAACL 2024, NAACL 2023, Springer Nature 2021, IEEE Access 2021, IEEE Access 2022, IEEE Access 2024

 TECHNICAL SKILLS

- **Programming:** Python, C language, Scala, Rust
- **Mobile and Web Technologies:** HTML, CSS, JavaScript
- **Cloud platforms:** Amazon web services, Google Cloud Platform, and Microsoft Azure
- **ML Tools:** PyTorch, Jax, Tensorflow, Keras, Scipy, Pandas, Numpy, DeepChem, LaTeX
- **DevOps and Workflow Tools:** Docker, MLflow

 TEACHING EXPERIENCE

- **Shala by IIT Bombay:** DL PI-2
Graph Convolutional Networks for NLP & Knowledge graphs

EXTRA ML PROJECTS

Covid-19 Question-Answering Bot [2020] Developed a Covid-19 question-answering pipeline that extracts keywords, retrieves and ranks relevant passages using vector search, and summarizes information using the BART model, deployed on telegram.

Image & Product Similarity in E-commerce [2018] Developed a product recommendation system that transforms product pages into graphs, applies graph isomorphism and image vector techniques to identify structural and visual similarities, enhancing recommendation accuracy

Music Generation with LSTM & Double Stacked GRU [2017] Developed a music generation system that transforms MIDI files into encoded matrices and trains single-layer and double-stacked layer models using LSTM and GRU architectures.



*equal contribution

Voice-Controlled Robotic Arm [2016] Developed a Raspberry Pi-controlled robotic arm with a text-to-speech module that translates vocal commands into diverse actions, securing second prize in a college technical exhibition for innovation.

 INVITED TALKS

- | | |
|--|-------------|
| Fine-Tuning Domain-Specific LLMs: When, Why, and How
<i>LinkedIn Office, India</i> | July, 2024 |
| LLMs in Healthcare & Lifescience domain
<i>Clinical NLP Group, KU Leuven University</i> | June, 2024 |
| Robust Evaluation of Medical LLMs: Current Issues and Future Directions
<i>Edinburgh Clinical NLP Group, The University of Edinburgh</i> | May, 2024 |
| OpenBioLLMs: Advancing Large Language Models in Medical Domain
<i>Rajpurkar Lab, Harvard Medical School</i> | May, 2024 |
| Hallucinations in LLMs: Causes, Types, and Mitigation Techniques , India
<i>ICCCSP conference & SSN College, Chennai</i> | March, 2024 |
| Adapting Large language models to low resource languages , Lucknow, India
<i>Google Developer Group, India</i> | Jan, 2024 |
| Parameter-Efficient Fine-Tuning with Low-Rank Adaptation , Kanpur, India
<i>Google Developer Group, DevFest India</i> | Dec, 2023 |
| Fine-Tuning Open-Source LLMs: Best Practices , Lucknow, India
<i>Google Developer Group, DevFest India</i> | Dec, 2023 |
| MLOps: The Keystone of Sustainable AI , Coimbatore, India
<i>Gradient Optimizers Meetup</i> | Jan, 2023 |
| Federated Learning & Distributional Shift in Healthcare , Chennai, India
<i>Gradient Optimizers Meetup</i> | Dec, 2022 |
| AI in Law: A New Legal Era , Kangra, India
<i>District Court Kangra</i> | Oct, 2021 |
| Reasoning in LLMs Through Math Word Problems , Chennai, India
<i>ML Researchers Meetup</i> | Oct, 2020 |
| Graphs Neural Networks for NLP , IITB, India
<i>Indian Institute of Technology Bombay, Shala</i> | Jul, 2020 |
| Functional Programming: Journey to the Decorator World , Manipal, India
<i>Manipal Institute of Technology, MUPy</i> | Oct, 2017 |
| A Deep Dive into IP Addresses , Lucknow, India
<i>Babu Banarasi Das University, Lucknow</i> | July, 2015 |

 FEATURED
OPEN-SOURCE
PROJECTS


- | | |
|---|--|
| LLMtuner
<i>Python</i> <ul style="list-style-type: none">• A module for Fine-Tune Llama, Whisper, and other LLMs with best practices like LoRA, QLoRA, through a sleek, scikit-learn-inspired interface | Nov, 2023
 (200+ stars) |
| Promptify
<i>Python and JavaScript</i> | Jan, 2023
 (3k+ stars) |

- A module for prompt engineering and versioning, Enabling users to efficiently utilize the GPT and similar prompt-based models to get structured output for various NLP tasks, including NER, QA, Classification, etc
- Github Trending repository

Research Papers Search (Resp)

Jul 15, 2022

Python

 (300+ stars)

- A module to Retrieves paper citations from Google Scholar
- Fetches relevant papers by keywords across sources like ACL, ACM, PMLR, etc.
- Extracts cough features including spectrograms, contiguous segments, and cough events, etc.
- Implements various ML and DL algorithms for respiratory audio analysis tasks including automated cough classification, clustering, anomaly detection, etc.

HONORS AND AWARDS

- Best NLP Researcher Oct, 2022
Saama Technologies, India
- Shining Star for the Month Award Nov, 2018
Saama Technologies, India
- 2nd prize in Technical and Robotics Exhibition Jun, 2015
Babu Banarasi Das University, Lucknow, India

POSITIONS OF RESPONSIBILITY

- Founder**, Open Life-Science AI Dec, 2023 - Present
 - Founded Open Life-Science AI, an open-source community dedicated to advancing Large Language Models (LLMs) development & integration in Healthcare.
- Community Lead**, Tensorflow Lucknow Group with Google Nov, 2023 - Present
 - Lead Google AI events for knowledge sharing and networking.
 - Guide workshops and discussions on TensorFlow/Jax/KerasNLP trends.
 - Develop tutorials and guides for TensorFlow/Jax application.
- Founder**, Lucknow AI Labs Oct, 2023 - Present
 - Spearheaded AI education programs in Tier 3 cities and villages across Uttar Pradesh for widespread AI literacy.
 - Mentored AI startups and developing AI solutions for local challenges
 - Working on building Multilingual large language and speech models for low-resource languages spoken in Uttar Pradesh, such as Awadhi and Magahi.
- Founder**, PromptsLab Dec, 2022 - Present
 - Founded PromptLab, an open-source community dedicated to advancing Large Language Models (LLMs) development & integration into robust NLP pipelines.
 - Developed open-source libraries like Promptify, and PromptifyJS to standardize workflow and reduce friction in consuming LLMs for production use cases.