

Thanmay Jayakumar

 /ThanmayJ |  /Thanmay |  ThanmayJ.GitHub.io |  thanmay2030@gmail.com

EDUCATION

INDIAN INSTITUTE OF TECHNOLOGY MADRAS
Chennai, India | 2024 - Present
(M.S.) Data Science & AI, NLP

VISVESVARAYA NATIONAL INSTITUTE OF TECHNOLOGY Nagpur, India | 2019 - 2023
(B.Tech) Electronics & Communication Engineering

COURSEWORK

- Generative AI with Large Language Models [\[Cert.\]](#)
- NVIDIA's Building Transformer-Based Natural Language Processing (NLP) Applications [\[Cert.\]](#)
- IIIT-Hyderabad's Summer School on NLP [\[Cert.\]](#)
- Stanford CS224n: NLP with DL [\[Course\]](#)
- IIT-Madras CS6910: Deep Learning [\[Course\]](#)
- NVIDIA's Fundamentals of Deep Learning [\[Cert.\]](#)
- Convolutional Neural Networks [\[Course\]](#)
- Data Structures and Algorithms [\[Cert.\]](#)
- Databases and SQL • Operating Systems
- Linear Algebra • Numerics and Probability

SKILLS

PROGRAMMING LANGUAGES

• Python • C • C++ • MATLAB • Perl (basic)

LIBRARIES

• PyTorch • Tensorflow • NumPy • SciPy • Pandas

SOFTWARE/TOOLS

• Git • Bash • MS Office • Adobe Photoshop
• LaTeX • HuggingFace • OS: Linux, Windows

LANGUAGES

Fluent	English, Tamil, Hindi, Telugu
Intermediate	German, Malayalam, Kannada
Elementary	Chinese, Sanskrit, Indonesian

EXTRACURRICULARS

- Workshop Coordinator:** Organized and taught various workshops on Data Science under the IEEE VNIT Student Chapter Nagpur.
- Volunteer:** Mentored junior year students at IvLabs on Deep Learning and NLP research.
- Graphic Designer:** Member of the Magazine & Literary Club, VNIT Nagpur.
- Piano & Music Theory:** Grade 5 - Associated Board of the Royal Schools of Music (ABRSM).

OPEN-SOURCE WORK

- Research Paper Notes [\[HackMD\]](#)
- Research Paper Implementations [\[GitHub\]](#)

EXPERIENCE

AI4BHARAT, IIT MADRAS

Sep 2023 - Present

AI Resident **Advisors:** Professors [R Dabre](#), [A Kunchukuttan](#), [M Khapra](#)

- Research on developing efficient multilingual LLMs.
- Experience with data collection, pretraining, instruction-tuning.
- [\[Publication\]](#) Released "Airavata", a Hindi instruction-tuned LLM.

IIT KANPUR

May - Aug 2022

Research Intern (SURGE 2022 Intern)

Advisor: Prof [Vipul Arora](#)

- Aimed at solving the task of Spoken Term Detection (STD) to retrieve queried speech files in an audio database.
- Implemented three different approaches to STD for query localization, classification and location suggestion in a database.
- Analyzed an optimal combination of the above, in order to work towards building a language-agnostic system.

PUBLICATIONS

- [\[Paper\]](#) JA Husain, R Dabre, A Kumar, J Gala, **Thanmay Jayakumar** et al. "RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization" (ACL 2024)
Awarded the ACL 2024 Senior Area Chair Award
- [\[Paper\]](#) A global team led by MBZUAI. "CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark" (NeurIPS 2024)
- [\[Paper\]](#) Fauzan Farooqui, **Thanmay Jayakumar**, Pulkit Mathur, Mansi Radke, "Leveraging Linguistically Enhanced Embeddings for Open Information Extraction" (LREC-COLING 2024)
- [\[Paper\]](#) **Thanmay Jayakumar**, Fauzan Farooqui, Luqman Farooqui, "Large Language Models are legal but they are not: Making the case for a powerful LegalLLM" (NLLP, EMNLP 2023)
- [\[Paper\]](#) Kshitij Ambilduke, **Thanmay Jayakumar**, Luqman Farooqui, Himanshu Padole, Anamika Singh, "Attending to Transforms: A Survey on Transformer based Image Captioning" (PCEMS 2023)

SELECTED PROJECTS

DEEP LEARNING NEURAL NETWORKS

- Distributed Processing and Sharding:** [\[GitHub\]](#)
Deployed PyTorch's Distributed Data Parallel and Fully Sharded Data Parallel with RoBERTa on a two-GPU parallelism.
- Neural Networks from Scratch:** [\[GitHub\]](#)
Implemented multiple neural networks from scratch using PyTorch and NumPy, including feedforward classifiers, convolutional neural networks, sequence models, and attention-based networks.

TEXT-BASED GENERATIVE MODELING

- Image Captioning:** [\[GitHub\]](#)
Surveyed image captioning methods for our bachelors thesis, and trained captioning models with a ResNet image encoder and various decoders in PyTorch using the Flickr caption dataset.
- Neural Machine Translation:** [\[GitHub\]](#)
Implemented Encoder-Decoder architectures from scratch in PyTorch using the Multi30k Dataset for German-English.