# Indraneil Paul

Email   Github   Scholar   Twitter   LinkedIn   Website

I am a doctoral researcher interested in optimising **code-generation LM pre-training**, emphasising **function calling** and **multilingual performance**, and contributing to multiple **open-source LM** releases. My mission is to unlock the application of LMs beyond text-only settings to areas like **robot navigation** and **agentic workflows** by improving their abilities to **reason**, **offload computation** and learn from **environment feedback**. I also work on **preference learning** methods to improve LMs' code generation capabilities along non-functional axes like **security** and **efficiency**. My interests span all facets of improving LM pre-training efficiency, including **data curation**, **context-length extension**, **modularity** and **sparse-expert models**.

## Education

| | | |
|---|---|---|
| 09/22 - Pres. | ELLIS PhD Candidate in Informatics, TU Darmstadt, Germany | Enrolment |
| 07/17 - 07/19 | Masters by Research in Computer Science, IIIT Hyderabad, India | Certificate |
| 08/13 - 05/17 | Bachelors of Technology in Computer Science, IIIT Hyderabad, India | Certificate |

## Summer Schools

| | | |
|---|---|---|
| 07/23 | Lisbon Machine Learning Summer School (LxMLS) | Certificate |
| 07/21 | European Summer School in Logic, Language and Information (ESSLLI) | Certificate |

## Invited Talks

| | | |
|---|---|---|
| 10/24 | Challenges in Code LMs, IIIT Hyderabad | Slides |
| 09/24 | Code Generation : Challenges and Solutions, BHT Berlin | Slides |
| 04/23 | Parameter-Efficient Fine-Tuning for NLP, MBZUAI | Slides |
| 01/23 | Multilingual Adapters, TU Darmstadt | Slides |

## Selected Publications

ObscuraCoder : Powering Efficient Code LM Pre-Training Via Obfuscation Grounding
ICLR 2025, Singapore — Abstract | PDF
**Indraneil Paul** et al.

BigCodeBench : Benchmarking Code Generation With Diverse Function Calls And Complex Instructions
ICLR 2025, Singapore — Abstract | PDF
Terry Yue Zhuo et al. (incl. **Indraneil Paul**)

EMMA-500 : Enhancing Massively Multilingual Adaptation Of Large Language Models
ACL 2025, Vienna (Under Review) — Abstract | PDF
Shaoxiong Ji et al. (incl. **Indraneil Paul**)

IRCoder : Intermediate Representations Make Language Models Robust Multilingual Code Generators
ACL 2024 Oral, Bangkok (Outstanding Paper) — Slides | Abstract | PDF
**Indraneil Paul** et al.

StarCoder 2 And The Stack V2 : The Next Generation
TMLR 2024 — Slides | Abstract | PDF
Anton Lozhkov et al. (incl. **Indraneil Paul**)

Adapters : A Unified Library For Parameter-Efficient And Modular Transfer Learning
EMNLP 2023 System Demonstrations, Singapore — Demo | Abstract | PDF
Clifton Poth et al. (incl. **Indraneil Paul**)

Sub-Task Imputation via Self-Labelling to Train Image Moderation Models on Sparse Noisy Data
CIKM 2022 Oral, Atlanta — Slides | Abstract | PDF
**Indraneil Paul** et al.

## Research Experience

09/22 - Pres. **Doctoral Researcher, TU Darmstadt Ubiquitous Knowledge Processing Lab, Darmstadt**

- Researching comparative benefits of various PEFT and MoE methods
- Implemented LLVM IR grounding for improving the multilingual performance of code LMs
- Demonstrated the benefits of pre-training code LMs with obfuscation grounding
- Investigating code LM improvement along non-functional axes like runtime
- Created and solely maintained **VLLM-Code-Harness**, a library for efficient code LM evaluation

`GPT-NeoX` `HuggingFace Transformers` `Axolotl` `TRL` `DistilLabel` `Python` `Docker` `LLVM`

06/17 - 08/19 **Research Assistant, IIIT-H Language Technologies Research Center, Hyderabad**

- Employed temporal activity, network and Tweet-based features to characterize verified users on Twitter
- Curated a **dataset** of 235K+ verified Twitter users, containing 79M+ edges and 494M+ Tweets

`Graph-Tool` `FastAI` `Neo4j` `AllenNLP` `Twitter API` `PoweRLaw` `Python` `R`

06/18 - 07/19 **Research Assistant, IIIT-H Machine Learning Lab, Hyderabad**

- Researched constraint-aware two-sided matching algorithms on dynamic bipartite graphs
- Benchmarked non-manipulable preference elicitation mechanisms for ride-sharing drivers

`ParamILS` `CVXOpt` `MATLAB` `Python` `C++`

## Industry Experience

04/20 - 08/22 **Applied Scientist, Amazon Inc. (Advertising), Bangalore**

- Created text, image and multi-modal models for improving EU ad moderation automation by 28%
- Researched multi-modal, multi-lingual and multi-task pre-training objectives for ad catalog tagging
- Devised sample-efficient training methods for ViT models using self-labelling and sub-task distillation

`HuggingFace Transformers` `PyTorch` `Python` `CUDA C++` `TensorRT` `AWS SageMaker`

07/19 - 03/20 **Software Development Engineer, Amazon Inc. (Logistics), Hyderabad**

- Implemented a planner enabling merchants to rank options and schedule last-mile package drop-offs
- Oversaw database tuning, JVM optimizations and message queue setup for event ingestion service

`Spring` `METIS` `Java` `AWS SNS` `AWS SQS` `AWS DynamoDB`

## Open Source Experience

04/24 - Pres. **MaLA-LM, UTTER Project**

- Conducted SOTA multilingual continual pre-training evaluations on frontier LMs
- Investigated the code completion performance of multilingual LMs in non-English language prompts
- Worked on the **EMMA-500** model and **MaLA-2** massively multilingual corpus releases

`HuggingFace Transformers` `Megatron-DeepSpeed` `DeepSpeed` `Python` `Docker`

06/23 - Pres. **BigCode Project, ServiceNow and HuggingFace**

- Contributed to **StarCoder-2** pre-training data collection and training ablations
- Worked on containerization, evaluation framework and annotation for **BigCodeBench**

`LLVM` `HuggingFace Transformers` `Megatron-LM` `Python` `Docker`

05/17 - 07/17 **Google Summer of Code, Green Navigation**

- Implemented an LSTM forecaster for the **EV-Charge-Prediction** project to alleviate range anxiety
- Implemented an ensemble solution that reduced absolute forecasting error by 39%
- Productionized the Bayesian Optimization service for optimal hyper-param selection in training jobs

`TensorFlow` `Pandas` `BayesOpt` `Python`

## References

| | | |
|---|---|---|
| TU Darmstadt | Prof. Dr. Iryna Gurevych, PhD Thesis Advisor | Email |
| JMU Wurzburg | Prof. Dr. Goran Glavas, PhD Thesis Co-Advisor | Email |
| IIIT Hyd. | Prof. Dr. Ponnurangam Kumaraguru, MSc Thesis Advisor | Email |