




# INDRANEIL PAUL



 Email  Github  HuggingFace  Scholar  Twitter  LinkedIn  Website

I am a doctoral researcher interested in optimising **code generation LM pre-training and post-training**, with an emphasis on **tool usage** and **multilingual performance**. My long-term mission is to enhance model tool use, unlocking the application of LMs beyond text-only settings to areas such as **robot navigation** and **agentic workflows** by improving their capabilities to **reason**, **offload computation**, and learn from **environmental feedback**. I also work on **preference learning** methods to enhance LMs' code generation capabilities along non-functional axes, such as **security** and **efficiency**. My interests span all facets of improving LM training efficacy, including **data curation**, **context length extension**, **modularity**, and **reinforcement learning**. I am also a long-standing open-source contributor to multiple public repositories and have worked on numerous **open-source LM** releases.

## 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

## INDUSTRY EXPERIENCE

10/25 - 03/26	Applied Scientist PhD Intern, Amazon Inc. (Web Services), Berlin <ul style="list-style-type: none"><li>➤ Researched RL methods to reduce the cloud tool-calling error-rate in <b>Amazon Q Developer</b> agent</li><li>➤ Explored asynchronous RL approaches for improved distributed training efficiency</li><li>➤ Automated RL environment creation for agents using infrastructure-as-code emulators</li></ul> <div><b>HuggingFace Transformers</b> <b>VeRL</b> <b>Megatron-LM</b> <b>LocalStack</b> <b>AWS Fargate</b> <b>AWS AgentCore</b> <b>AWS SageMaker</b></div>
04/20 - 08/22	Applied Scientist, Amazon Inc. (Advertising), Bangalore <ul style="list-style-type: none"><li>➤ Created text, image and multi-modal models for improving EU ad moderation automation by 28%</li><li>➤ Researched multi-modal, multi-lingual and multi-task pre-training objectives for ad catalog tagging</li><li>➤ Devised sample-efficient training methods for ViT models using self-labelling and sub-task distillation</li></ul> <div><b>HuggingFace Transformers</b> <b>PyTorch</b> <b>CUDA C++</b> <b>TensorRT</b> <b>AWS SageMaker</b></div>
07/19 - 03/20	Software Development Engineer, Amazon Inc. (Logistics), Hyderabad <ul style="list-style-type: none"><li>➤ Implemented a planner enabling merchants to rank options and schedule last-mile package drop-offs</li><li>➤ Oversaw database tuning, JVM optimizations and message queue setup for event ingestion service</li></ul> <div><b>Spring</b> <b>METIS</b> <b>Java</b> <b>AWS SNS</b> <b>AWS SQS</b> <b>AWS DynamoDB</b></div>





## RESEARCH EXPERIENCE

09/22 - 12/26	Doctoral Researcher, TU Darmstadt Ubiquitous Knowledge Processing Lab, Darmstadt <ul style="list-style-type: none"><li>➤ Researched comparative benefits of various PEFT and MoE methods</li><li>➤ Implemented LLVM IR grounding for improving the multilingual performance of code LMs</li><li>➤ Demonstrated the benefits of pre-training code LMs with obfuscation grounding</li><li>➤ Investigating code LM improvement along non-functional axes like runtime</li><li>➤ Created and solely maintained <b>VLLM-Code-Harness</b>, a library for efficient code LM evaluation</li></ul> <div><b>GPT-NeoX</b> <b>HuggingFace Transformers</b> <b>Axolotl</b> <b>TRL</b> <b>DistilLabel</b> <b>Docker</b> <b>LLVM</b></div>
06/17 - 08/19	Research Assistant, IIIT-H Language Technologies Research Center, Hyderabad <ul style="list-style-type: none"><li>➤ Employed temporal activity, network and Tweet-based features to characterize verified users on Twitter</li><li>➤ Curated a <b>dataset</b> of 235K+ verified Twitter users, containing 79M+ edges and 494M+ Tweets</li></ul> <div><b>Graph-Tool</b> <b>FastAI</b> <b>Neo4j</b> <b>AllenNLP</b> <b>Twitter API</b> <b>PowerLaw</b> <b>R</b></div>
06/18 - 07/19	Research Assistant, IIIT-H Machine Learning Lab, Hyderabad <ul style="list-style-type: none"><li>➤ Researched constraint-aware two-sided matching algorithms on dynamic bipartite graphs</li><li>➤ Benchmarked non-manipulable preference elicitation mechanisms for ride-sharing drivers</li></ul> <div><b>ParamLLS</b> <b>CVXOpt</b> <b>MATLAB</b> <b>C++</b></div>

## OPEN SOURCE EXPERIENCE

- 04/24 - 09/24 **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** **Docker**
- 06/23 - 09/24 **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** **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**

## 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

### DROID : A RESOURCE SUITE FOR AI-GENERATED CODE DETECTION

EMNLP 2025, Suzhou

Daniil Orel et al. (incl. Indraneil Paul)

 **ABSTRACT** |  **PDF**

### EMMA-500 : ENHANCING MASSIVELY MULTILINGUAL ADAPTATION OF LARGE LANGUAGE MODELS

DMLR 2026 (Under Review)

Shaoxiong Ji et al. (incl. Indraneil Paul)

 **ABSTRACT** |  **PDF**

### OBSCURACODER : POWERING EFFICIENT CODE LM PRE-TRAINING VIA OBFUSCATION GROUNDING

ICLR 2025 Poster, Singapore

Indraneil Paul et al.

 **ABSTRACT** |  **PDF**


### BIGCODEBENCH : BENCHMARKING CODE GENERATION WITH DIVERSE FUNCTION CALLS AND COMPLEX INSTRUCTIONS

ICLR 2025 Oral, Singapore

Terry Yue Zhuo et al. (incl. Indraneil Paul)

 **SLIDES** |  **ABSTRACT** |  **PDF**

### IRCODER : INTERMEDIATE REPRESENTATIONS MAKE LANGUAGE MODELS ROBUST MULTILINGUAL CODE GENERATORS

ACL 2024 Oral, Bangkok ( **Outstanding Paper**)

Indraneil Paul et al.

 **SLIDES** |  **ABSTRACT** |  **PDF**

### STARCORDER 2 AND THE STACK V2 : THE NEXT GENERATION

TMLR 2024

Anton Lozhkov et al. (incl. Indraneil Paul)

 **SLIDES** |  **ABSTRACT** |  **PDF**

### ADAPTERS : A UNIFIED LIBRARY FOR PARAMETER-EFFICIENT AND MODULAR TRANSFER LEARNING

EMNLP 2023 System Demonstrations, Singapore

Clifton Poth et al. (incl. Indraneil Paul)

 **DEMO** |  **ABSTRACT** |  **PDF**

### SUB-TASK IMPUTATION VIA SELF-LABELLING TO TRAIN IMAGE MODERATION MODELS ON SPARSE NOISY DATA

CIKM 2022 Oral, Atlanta

Indraneil Paul et al.

 **SLIDES** |  **ABSTRACT** |  **PDF**