

INDRANEIL PAUL

 Email  Github  HuggingFace  Scholar  Twitter  LinkedIn  Website

I am a researcher interested in optimizing the **mid-training and post-training** of Language Models (LMs), with an emphasis on **agentic coding abilities** and **tool use**. My long-term mission is to enhance LMs' **long-horizon** operation, unlocking their application beyond conventional settings to areas such as **computer use** and **recursive workflows** by improving their capabilities to **reason**, **offload computation**, and learn from **environmental feedback**. I also work on **preference learning** and **verifiers**, aiming to enhance LMs' capabilities beyond 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**.

EDUCATION

09/22 – 10/26	ELLIS PhD Candidate in Informatics, TU Darmstadt, Germany	 ENROLMENT
07/17 – 07/19	Master's by Research in Computer Science, IIIT Hyderabad, India	 CERTIFICATE
08/13 – 05/17	Bachelor of Technology in Computer Science, IIIT Hyderabad, India	 CERTIFICATE

INDUSTRY EXPERIENCE

10/25 – 04/26	Applied Scientist PhD Intern, Amazon Inc. (Web Services), Berlin ➤ Researched RL methods to reduce the cloud tool-calling error rate in the Amazon Q Developer agent ➤ Explored asynchronous RL approaches for improved distributed training efficiency ➤ Automated RL environment creation for agents using infrastructure-as-code emulators HuggingFace TransformersVeRLMegatron-LMLocalStackAWS FargateAWS AgentCoreAWS SageMaker
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-labeling and sub-task distillation HuggingFace TransformersPyTorchCUDA C++TensorRTAWS 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 an event ingestion service SpringMETISJavaAWS SNSAWS SQSAWS DynamoDB

CONTRIBUTOR 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 TransformersMegatron-DeepSpeedDeepSpeedDocker
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 LLVMHuggingFace TransformersMegatron-LMDocker
05/17 – 07/17	Google Summer of Code, Green Navigation ➤ Implemented an LSTM forecaster for the EV-Charge-Prediction project to alleviate range anxiety ➤ Developed an ensemble solution that reduced absolute forecasting error by 39% ➤ Productionized the Bayesian Optimization service for optimal hyperparameter selection in training jobs TensorFlowPandasBayesOpt

SUMMER SCHOOLS

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

ACADEMIC SERVICE

07/26	Organizer, SemEval 2026 Task on GenAI Code Detection & Attribution	 EVENT PAGE
08/22	Program Committee, KDD 2022 Content Generation for e-Commerce Workshop	 EVENT PAGE

 SELECTED PUBLICATIONS

OctoLong : Mid-Training on Cross-Repository Code Contexts Enhances Long-Context Modeling
NAACL 2027 (Under Review)
Indraneil Paul et al.

 [ABSTRACT](#) |  [PDF](#)

Themis : Training Robust Multilingual Code Reward Models for Flexible Multi-Criteria Scoring
TMLR 2026 (Under Review)
Indraneil Paul et al.

 [ABSTRACT](#) |  [PDF](#)

Aletheia : What Makes RLVR for Code Verifiers Tick?
TMLR 2026
Vatsal Venkatkrishna et al. (incl. Indraneil Paul)

 [ABSTRACT](#) |  [PDF](#)

AICD Bench : A Challenging Benchmark for AI-Generated Code Detection
EACL 2026, Rabat
Daniil Orel et al. (incl. Indraneil Paul)

 [SLIDES](#) |  [ABSTRACT](#) |  [PDF](#)

DROID : A Resource Suite for AI-Generated Code Detection
EMNLP 2025, Suzhou
Daniil Orel et al. (incl. Indraneil Paul)

 [SLIDES](#) |  [ABSTRACT](#) |  [PDF](#)

ObscuraCoder : Powering Efficient Code LM Pre-Training via Obfuscation Grounding
ICLR 2025 Poster, Singapore
Indraneil Paul et al.

 [SLIDES](#) |  [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](#)

StarCoder 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-Labeling to Train Image Moderation Models on Sparse Noisy Data
CIKM 2022 Oral, Atlanta
Indraneil Paul et al.

 [SLIDES](#) |  [ABSTRACT](#) |  [PDF](#)

 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

 REFERENCES

AWS Berlin	Dr. György Szarvas, Research Internship Advisor	 EMAIL
TU Darmstadt	Prof. Dr. Iryna Gurevych, PhD Thesis Advisor	 EMAIL
JMU Würzburg	Prof. Dr. Goran Glavaš, PhD Thesis Advisor	 EMAIL
IIIT Hyd.	Prof. Dr. Ponnurangam Kumaraguru, MSc Thesis Advisor	 EMAIL