

Atharva Deshmukh

B.Tech, Computer Science & Engineering
Indian Institute of Technology, Patna

Email: atharva1763@gmail.com

LinkedIn: [Atharva Deshmukh](#)

GitHub: [github @ Atharva1763](#)

EDUCATION

• Indian Institute of Technology (IIT), Patna

Bachelor of Technology in Computer Science and Engineering (2021 – 2025)

Bihar, India

Grades: 89.8%

PUBLICATIONS

• XWind: A Cross-site Router for Large Language Model Inference Serving at Renewable Energy Farms

ASPLOS 2027

Authors – Tella Rajashekhar Reddy, [Atharva Deshmukh](#), Liangcheng Yu, Chaojie Zhang, Mike Shepperd, Rohan Gandhi, Anjaly Parayil, Srinivasan Iyengar, Ajay Manchepalli, Debopam Bhattacharjee

Under Review, Paper

• BeLLMan: Controlling LLM Congestion

FAISys 2025

Authors – Tella Rajashekhar Reddy*, [Atharva Deshmukh](#)*, Karan Tandon, Rohan Gandhi, Anjaly Parayil, Debopam Bhattacharjee

Paper

• Can Language Models Persuade? Exploring the Persuasive Efficacy of Large Language and Vision Language Models

Computer Speech & Language

Authors – Rohan Kirti, [Atharva Deshmukh](#), Kiran Kumar Dugana, Yash Rathore, Shipra Shriparn, Sriparna Saha, Roshni R Ramnani, Anutosh Maitra

Paper

• Dynamic Negotiation Landscapes: MBPS and the Interplay of Buyer Personalities

Neurocomputing

Authors – Subrata Das*, [Atharva Deshmukh](#)*, Sriparna Saha, Roshni Ramnani, Anutosh Maitra, Shubhashis Sengupta

Under Review

• APTSumm at BioLaySumm Task 1: Biomedical Breakdown, Improving Readability by Relevancy-Based Selection

BioNLP Workshop, ACL 2023

Authors – A.S. Poornash, [Atharva Deshmukh](#), Archit Sharma, Sriparna Saha (* = Equal Contribution)

Paper

EXPERIENCE

• Optimizing LLM Inference under Congestion @ [Microsoft Research](#)

Jan. 2025 – Present

Position: Research Intern | On-Site, Bengaluru

- Optimized LLM inference **under congestion** via token-level output reduction, improving throughput by **19%** and reducing energy use by **25%**; built predictive scheduling models for output-length estimation, ensuring fairness and efficiency in vLLM serving during peak load.
- Designed a **congestion-control mechanism** for query scheduling and GPU resource allocation across inference clusters.
- Building **Heron** – a power-, hardware-, and workload-aware routing system for renewable-powered AI inference clusters to optimize performance under variable energy supply.
- Work is **proprietary**; focuses on optimizing large-scale inference workloads.

• LLM Training Optimization @ [IBM Research](#)

May 2024 – Aug. 2024

Position: Hybrid Cloud Intern | On-Site, Bangalore

- Designed a novel framework for loss-curve analysis, leveraging **algorithmic** and **statistical** signals to recommend optimal parameter configurations; analyzed training data from 100+ epoch-trained LLMs.
- Implemented a modular codebase, orchestrating experiments to predict optimal model configurations for training – currently used internally.
- Built **early-stopping techniques** that achieved roughly **20% reduction in training epochs** while maintaining model performance.

• Multi-Agent RL for Personality-Aware Negotiation (MBPS) @ AI-NLP Lab, IIT Patna

Aug. 2023 – Feb. 2024

Guide: Dr. Sriparna Saha | Sponsored project with Accenture Labs

Under Review

- Co-designed the experimental protocol, tuned the stateless/stateful UCB reward functions, and ran seller-model ablations across all four buyer personalities (Aggressive, Negotiator, Avoider, Submissive).
- Devised a **multi-agent RL** strategy with a **UCB1**-based seller agent over the new **MBPS** simulator, adapting bargaining behaviour to diverse buyer personalities; work under review at *Neurocomputing*.
- **LLM Adaptation & High-Throughput Inference @ CNeRG Lab,** May 2023 – Aug. 2023
IIT Kharagpur
Position: NLP Research Intern (Full-time) / Remote
 - Built expertise in **HTCondor** for efficient GPU job scheduling; worked with **LangChain**, **LlamaIndex**, and training/inference stacks for **Falcon**, **MPT**, and **Llama**.
 - Implemented SOTA LLM training techniques including **QLoRA**, **LoRA**, and **adapter tuning**.
- **Lay Summarization for Biomedical Articles (APTSumm) @** Dec. 2022 – April 2023
AI-NLP Lab, IIT Patna
Guide: Dr. Sriparna Saha / BioLaySumm Shared Task, ACL 2023 Paper
 - Built a **three-step abstractive** summarization pipeline that sections the document, generates candidate summaries via **SimCLS**, then re-ranks and selects the top-performing paragraph per section.
 - Achieved **2nd rank** on readability scores and strong ROUGE numbers on the BioLaySum shared task; co-authored the resulting paper at the **ACL 2023 BioNLP Workshop**.

PROJECTS

- **API-Driven Multi-Step Query Resolution Chatbot @ DevRev,** Oct. 2023 – Dec. 2023
Inter-IIT Tech Meet
Team Lead Code
 - **Led** a team to build a chatbot for **multi-step API-orchestrated queries** – supporting add/delete/update of APIs and detecting unanswerable queries.
 - Three-stage pipeline using **LangChain**, **ChromaDB**, and **HuggingFace**: multi-query retrieval + reranking, output generation, and a feedback loop with dynamic few-shot examples.
- **Domain-Specific Extractive Question Answering @ DevRev,** Dec. 2022 – Feb. 2023
Inter-IIT Tech Meet
Bagged 1st position among all IITs Code
 - Built domain-adaptable rankers fine-tuned via **knowledge distillation** to re-rank **BM25**-retrieved passages; our model outperformed all other teams' approaches.
 - Combined retriever, ranker, and reader signals into an answerability classifier reaching up to **97% accuracy**.
 - Trained multiple reader models and developed data-augmentation strategies + evaluation scripts.

SCHOLASTIC ACHIEVEMENTS

- **Amazon ML Challenge '24:** Secured **3rd Rank** among **1000+** teams in an ML-based hackathon Certificate
- **JLR GearShift Code Rally '23:** Secured **All India Rank 4**, with participation from all top IITs Certificate
- **Inter IIT Tech Meet 11.0 '23:** Bagged **1st position** among all IITs in an NLP-based high-prep event (DevRev – Domain-Specific Question Answering)
- **K.V. Chess Nationals '17:** Secured **17th Rank** representing the state of Maharashtra Certificate
- **JEE Advanced '21:** Secured All India Rank **2644** out of 250,000+ candidates nationwide
- **JEE Mains '21:** Secured All India Rank **2557** out of 1.1 million candidates from across the country
- **Rupin Pass Trek '25:** Completed a high-altitude Himalayan trek (~15,000 ft) Photo

POSITIONS OF RESPONSIBILITY & EXTRA-CURRICULAR ACTIVITIES

- **Founder, Research Club, IIT Patna** (2023 – 2025): Established the club; organised guest talks and poster sessions to seed a research-oriented culture on campus.
- **Head, Machine Learning Club, NJack** (IIT Patna CS Club): Oversaw activities and built a learning community for ML enthusiasts – paper readings, workshops, and project sprints.
- **NSS** – volunteered in community service initiatives and outreach projects.