

✉ aflahkhan.2025@gmail.com

🌐 aflah02.github.io

in mohammad-aflah-khan

🔗 aflah02

🔑 IM\_mP38AAAAJ

# Mohammad Aflah Khan

## Education

- 2020-2024 **B.Tech. in Computer Science and Engineering**, *Indraprastha Institute of Information Technology, Delhi*, CGPA: 9.63/10  
Ranked 3rd in my graduating class
- 2018-2020 **CBSE - Grade XII (Science)**, *Lal Bahadur Shastri School*, 95%
- 2006-2018 **CBSE - Grade X**, *Banyan Tree School*, 95.8%

## Experience

- Aug 2023 – Present **Research Software Engineer**, *Max Planck Institute for Software Systems (MPI-SWS)*, Remote
- Working with Dr. Krishna P. Gummadi on research and engineering for large language models (LLMs). I joined as an intern in August 2023, transitioned to part-time in October 2023, and moved to a full-time role in June 2024 after graduating.
- Conducted research on:
    - Optimizing LLM pre-training and inference pipelines;
    - Building LLM testbeds to study memorization risks;
    - Studying LLM memorization dynamics and the effects of Parameter-Efficient Fine-Tuning (PEFT);
    - Investigating knowledge acquisition and factual correctness in LLMs.
  - Designed, built, and currently maintain internal tools:
    - **OpenChat** — internal chatbot interface;
    - **MaxCast** — research paper-to-podcast generation service;
    - **MaxChat** — document-grounded chat system.
  - Implemented these systems end-to-end, including on-premise model hosting and fine-tuning for performance.
  - Published and submitted research to top-tier (A\*) venues.
- Dec 2022 – Present **Community Researcher**, *EleutherAI*, Remote
- Contributing to open-source research on multilingual instruction tuning, model evaluation, and LLM memorization.
- Currently working on building a large-scale instruction-tuning corpus for Hindi.
  - Contributed to *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling* (ICML'23): Made major contributions to gender-bias evaluation and intervention case study. The models have been downloaded over 18 million times as of April 2025.
  - Contributed to *Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon* (ICLR'25): Developed an intuitive taxonomy for memorized sequence types and built corresponding predictors based on these memorization classes.

June 2021 – **Undergraduate Student Researcher**, *Laboratory for Computational Social Systems (LCS2)*, IIIT Delhi, India

Worked on computational social science and NLP research spanning hate speech, safety, and LLM evaluation.

- Worked on hate speech normalization (KDD'22) and designed recommendation strategies for fine-tuning improved hate-speech detectors (EACL'24).
- Led the **QUENCH** project (COLING'25), a benchmark for evaluating advanced reasoning abilities in large language models, with emphasis on Indic linguistic and cultural contexts.
- Part of the organizing team for **HASOC'23 Task 3**, focused on hateful span detection.
- Contributed to multiple projects involving dataset creation, model evaluation, and analysis of social biases in NLP systems.

May 2023 – **Summer Analyst**, *Goldman Sachs*, India

July 2023 Worked in the Finance, Planning & Analysis Engineering division on internal tools and process improvements.

- Revamped the central hub of the department to streamline workflows and information access.
- Built proof-of-concepts (POCs) based on user feedback to enhance search and access experience on the internal web application.
- Received a return offer to join full-time as an Analyst.

May 2022 – **Open Source Developer**, *Google Summer of Code – TensorFlow*, Remote

Sept 2022 Contributed to KerasNLP within the Keras/TensorFlow ecosystem, focusing on industry-oriented NLP solutions.

- Added support for data augmentation layers to KerasNLP, collaborating with Matthew Watson and Chen Qian.
- Contributed bug fixes and improvements to tokenizers and transformer encoder & decoder modules.
- Enhanced utilities to improve the usability and performance of KerasNLP for practical NLP workflows.

### Miscellaneous

Jan 2023 – **Teaching Assistant - Machine Learning**, *Computer Science and Engineering (CSE) Department of IIIT Delhi*, New Delhi

May 2023

As a Teaching Assistant under Dr. Anubha Gupta for the Machine Learning elective course with nearly 70 students, I conducted coding tutorials to help students get comfortable with the required libraries and tools. I also assisted in preparing exam questions, distributing reading materials, and grading assignments, while maintaining regular office hours to provide ongoing support and address student queries throughout the course.

May 2022 – **Teaching Assistant - Data Structures & Algorithms**, *Computer Science and Engineering (CSE) Department of IIIT Delhi*, New Delhi

Aug 2022

As a Teaching Assistant under Dr. Piyush Kedia for the core course Data Structures and Algorithms, which had nearly 600 enrolled students, I conducted weekly tutorials and lab sessions to reinforce key concepts and guide students through hands-on problem-solving. Alongside holding regular office hours, I frequently organized extra sessions to address doubts, provide individualized support, and ensure that students remained on track throughout the course.

---

## Publications

\* Implies Equal Contribution

- 2026 **Hubble: a Model Suite to Advance the Study of LLM Memorization** — Johnny Tian-Zheng Wei\*, Ameya Godbole\*, **Mohammad Aflah Khan\***, Ryan Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, Robin Jia; Oral @ ICLR 2026 | DATA-FM Workshop @ ICLR 2026 (Non-Archival).
- 2026 **In Agents We Trust, but Who Do Agents Trust? Latent Source Preferences Steer LLM Generations** — **Mohammad Aflah Khan**, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna P. Gummadi, Manish Gupta, Abhilasha Ravichander; ICLR 2026 | IASEAI 2026 (Non-Archival) | R2-FM Workshop @ ICML 2025 (Non-Archival).
- 2026 **Rote Learning Considered Useful: Generalizing over Memorized Data in LLMs** — Qinyuan Wu, Soumi Das, Mahsa Amani, Bishwamittra Ghosh, **Mohammad Aflah Khan**, Krishna P. Gummadi, Muhammad Bilal Zafar; ICLR 2026 | IASEAI 2026 (Non-Archival) | MemFM Workshop @ ICML 2025 (Non-Archival).
- 2026 **Fine-tuning vs. In-context Learning in Large Language Models: A Formal Language Learning Perspective** — Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, **Mohammad Aflah Khan**, Deepak Garg, Krishna P. Gummadi, Evimaria Terzi; Oral @ ACL 2026; Senior Area Chair Highlight Award
- 2026 **Position: Don't Just "Fix it in Post": A Science of AI Must Study Training Dynamics** — Stella Biderman, **Mohammad Aflah Khan**, Niloofar Miresghallah, Catherine Arnett, Fazl Barez, Naomi Saphra; Oral @ ICML 2026 Position Paper Track
- 2026 **Fractional Rotation, Full Potential? Investigating Performance and Convergence of Partial RoPE** — **Mohammad Aflah Khan**, Krishna P. Gummadi, Manish Gupta, Abhilasha Ravichander; EMNLP 2025.
- 2026 **Revisiting Privacy, Utility, and Efficiency Trade-offs when Fine-Tuning Large Language Models** — Soumi Das, Camila Kolling, **Mohammad Aflah Khan**, Mahsa Amani, Bishwamittra Ghosh, Qinyuan Wu, Till Speicher, Krishna P. Gummadi; IASEAI 2026 (Non-Archival).
- 2025 **TokenSmith: Streamlining Data Editing, Search, and Inspection for Large-Scale Language Model Training and Interpretability** — **Mohammad Aflah Khan\***, Ameya Godbole\*, Johnny Tian-Zheng Wei, Ryan Wang, James Flemings, Krishna Gummadi, Willie Neiswanger, Robin Jia; EMNLP 2025 System Demonstrations.
- 2025 **Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon** — USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, **Mohammad Aflah Khan**, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, Naomi Saphra; ICLR 2025.

- 2025 **Towards Reliable Latent Knowledge Estimation in LLMs: Zero-Prompt Many-Shot Based Factual Knowledge Extraction** — Qinyuan Wu, *Mohammad Aflah Khan*, Soumi Das, Vedant Nanda, Bishwamittra Ghosh, Camila Kolling, Till Speicher, Laurent Bindschaedler, Krishna P Gummadi, Evimaria Terzi; WSDM 2025.
- 2025 **QUENCH: Measuring the gap between Indic and Non-Indic Contextual General Reasoning in LLMs** — *Mohammad Aflah Khan\**, Neemesh Yadav\*, Sarah Masud, Md Shad Akhtar; COLING 2025.
- 2025 **Rethinking Memorization Measures in LLMs: Recollection vs. Counterfactual vs. Contextual Memorization** — Bishwamittra Ghosh, Soumi Das, Qinyuan Wu, *Mohammad Aflah Khan*, Krishna P. Gummadi, Evimaria Terzi, Deepak Garg; MemFM Workshop @ ICML 2025.
- 2024 **Understanding the Mechanics and Dynamics of Memorisation in Large Language Models: A Case Study with Random Strings** — Till Speicher, *Mohammad Aflah Khan*, Qinyuan Wu, Vedant Nanda, Soumi Das, Bishwamittra Ghosh, Krishna P. Gummadi, Evimaria Terzi.
- 2024 **The Duality of Hope: A Critical Examination of Controversial Annotations in HopeEDI** — *Mohammad Aflah Khan\**, Neemesh Yadav\*, Diksha Sethi\*, Raghav Sahni\*; ICLR 2024, Tiny Papers Track.
- 2024 **Probing Critical Learning Dynamics of PLMs for Hate Speech Detection** — Sarah Masud\*, *Mohammad Aflah Khan\**, Vikram Goyal, Md Shad Akhtar, Tanmoy Chakraborty; EACL Findings 2024.
- 2023 **Overview of the HASOC Subtracks at FIRE 2023: Detection of Hate Spans and Conversational Hate-Speech** — Shrey Satapara, Sarah Masud, Hiren Madhu, *Mohammad Aflah Khan*, Md Shad Akhtar, Tanmoy Chakraborty, Sandip Modha, Thomas Mandl; FIRE 2023.
- 2023 **Overview of the HASOC Subtrack at FIRE 2023: Identification of Tokens Contributing to Explicit Hate in English by Span Detection** — Sarah Masud, *Mohammad Aflah Khan*, Md Shad Akhtar, Tanmoy Chakraborty; Working Notes of FIRE 2023.
- 2023 **Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling** — Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, *Mohammad Aflah Khan*, et al.; ICML 2023.
- 2023 **The Art of Embedding Fusion: Optimizing Hate Speech Detection** — *Mohammad Aflah Khan\**, Neemesh Yadav\*, Mohit Jain, Sanyam Goyal; ICLR 2023, Tiny Papers Track.
- 2023 **Beyond Negativity: Re-Analysis and Follow-Up Experiments on Hope Speech Detection** — Neemesh Yadav\*, *Mohammad Aflah Khan\**, Diksha Sethi, Raghav Sahni; ICLR 2023, Tiny Papers Track.
- 2022 **Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization** — Sarah Masud, Manjot Bedi, *Mohammad Aflah Khan*, Md Shad Akhtar, Tanmoy Chakraborty; KDD 2022.

---

## Talks

- 2025 *LLMs at Scale* — Max Planck Computing and Data Facility (MPCDF), AI Kick-off Workshop, April 2025
- 2025 *LLMs at Scale* — Max Planck Institute for the Science of Light, Hosted by Florian Marquardt, April 2025.
- 2025 *An Overview of DeepSeek-V3/R1* — Max Planck Institute for Software Systems (MPI-SWS), AI, Computing & Society Initiative, February 2025.
- 2024 *Democratizing and Accelerating Research with LLMs: Making Science More Accessible Whilst Finding Interesting Research Problems* — Max Planck Institute for Security and Privacy (MPI-SP), Hosted by Meeyoung Cha, December 2024.
- 2024 [Demo + Lightning Talk] *Empowering Research with Open-Access LLMs: From Tools to Copilots* — AI, Computing & Society Initiative Launch Event at Max Planck Institute for Software Systems (MPI-SWS), December 2024.
- 2023 *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling* — Max Planck Institute for Software Systems (MPI-SWS), Hosted by Krishna Gummadi, July 2023
- 2023 *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling* — Goldman Sachs, Internal NLP/IR Reading Group, June 2023.

## Awards and Achievements

- 2022-23 Dean's List for Innovation in Research and Development — Recognized for outstanding contributions to research and development.
- 2021-22, 2022-23 Dean's List for Academic Excellence — Awarded for consistent high academic performance.
- 2022 Amazon ML Summer School — Selected to participate in advanced machine learning training.
- 2022 ACL Year-Round Mentorship Program — Selected for mentorship in natural language processing research.
- 2021 Anveshan Hackathon Finalist — Designed a Mindmap Generator from YouTube transcripts using open-source models.
- 2021 Byld + WiT Hackathon Runner Up — Created a Discord Bot that converts toxic messages into humorous alternatives.
- 2021 ACM-IIITD Induction Ideathon Runner Up — Designed a social interface to keep students updated during the pandemic.
- 2020 JEE Advanced — Qualified with All India Rank 5212.
- 2020 JEE Mains — Paper 1: 99.34 percentile; Paper 2: All India Rank 491 (Top 0.3%).
- 2020 UGEE — Attained All India Rank 130 in Undergraduate Entrance Examination (UGEE).
- 2020 MET — Attained All India Rank 265 in Manipal Entrance Test (MET).
- 2020 VITEEE — Attained All India Rank 1809 in Vellore Institute of Technology Engineering Entrance Examination (VITEEE).

---

## Organizer, Reviewer & Volunteer Work

- 2025 Reviewer for ICLR
- 2025 Program Committee Member - The First Workshop on Large Language Model Memorization (L2M2) @ ACL'25
- 2024 onwards Reviewer for ACL Rolling Review
  - 2023 Organizer for HateNorm 2023 — Shared Task on Identification of Tokens Contributing to Explicit Hate in Text under HASOC 2023.
  - 2024 Reviewer for International Conference on Computational Linguistics (COLING).
  - 2024 Reviewer for The Technical Symposium on Computer Science Education (SIGCSE TS).
- 2023 & 2024 Reviewer for The Workshop on Online Abuse and Harms (WOAH) (Colocated with \*CL conferences).
- 2023 Reviewer for the 11th International Conference on Big Data and Artificial Intelligence (BDA).
- 2022 Volunteer at ICON 2022 Conference

---

## Skills

- Relevant Coursework Natural Language Processing, Deep Learning, Machine Learning, Information Retrieval, Reinforcement Learning, Data Science, Linear Algebra, Probability & Statistics, Multivariate Calculus.
- Programming Python, Java, HTML, CSS, JavaScript, Kotlin.
- Libraries & Frameworks Huggingface, Deepspeed, PyTorch, TensorFlow, Keras, Streamlit, Scikit-Learn, FastAPI, React, TailwindCSS, Git, Linux, Bash Scripting, Slurm.