Rajat Kulshreshtha

LinkedIn Google Scholar Website rk.kuls@gmail.com +1-732-397-3890

Education

Carnegie Mellon University Pittsburgh, PA
Master of Language Technologies (Research Master's), Language Technologies Institute 2018
Focus areas: multilinguality, speech recognition, representation learning, NLP

Indian Institute of Technology, Guwahati Guwahati, India
B. Tech., Electronics & Electrical Engineering (with a Minor in Computer Science) 2015

Work Experience

Amazon AGI Foundations Pittsburgh, PA
Applied Scientist Aug 2022 – Present

- Core contributor to the Nova 2 multimodal LLM family (Lite, Pro, Omni, Sonic), focusing on inference-time optimization, speculative decoding, and E2E latency/throughput across modalities.
  - Led speculative decoding systems for large-scale LLM inference, achieving up to ~4–5× end-to-end speedups while preserving output quality and model alignment
  - Drove TTFT and tokens-per-second improvements through architectural and systems-level optimizations; work shipped in production serving real-time user traffic
- Designed and productionized evaluation, benchmarking, and release pipelines for large and on-device language models, enabling repeatable latency, throughput, and quality comparisons across model variants and hardware targets.

Telling.ai Pittsburgh, PA
Founding Engineer - CMU Startup Oct 2018 – Jul 2022

- Led development of cloud-based, speech-first AI systems for longitudinal monitoring of respiratory diseases such as COPD using voice and breath analysis.
- Built and maintained production-grade pipelines for data preprocessing, model training, and real-time inference, ensuring safety and reliability for clinical use.
- Developed models, published peer-reviewed studies, and co-invented patents that were crucial in securing an NSF SBIR grant and series A funding.

Goldman Sachs Bangalore, India
Senior Analyst, Securities Division Jun – Nov 2015

- Trained machine-learning models for automated order surveillance using Monte Carlo–simulated order flow scenarios derived from historical exchange data.

Marooner Technologies Guwahati, India
Cofounder - IIT Guwahati startup Apr 2014 - May 2015

- Developed a virtual trial room software that allows users to try on apparel by using novel image processing algorithms to generate target apparel images on the user.

Research Experience

Carnegie Mellon University Aug 2016 – Aug 2018
Graduate Research Assistant – Advisors: Prof. Bhiksha Raj, Prof. Florian Metze

- Conducted research on adversarial robustness in ASR systems, designing and evaluating targeted perturbations to expose model vulnerabilities and inform safer deployment strategies.

- Developed adaptive speech and conversation models to study early language acquisition in infants.

Research Intern – Advisors: Prof. Bhiksha Raj, Prof. Rita Singh Dec 2013; May – Jul 2014

- Built an information mining system for Polly, a voice-based service reaching low-literate populations in India, delivering developmental content through simple telephonic interactions.
- Analyzed large-scale call data to identify usage patterns, track network growth, and inform content design to maximize impact.
- Created a novel method using Bayesian inference and semantic analysis to find the safest travel route based on modeling of crime distribution from police records and news feeds.

Hanyang University, South Korea *May – Jul 2013*
Research Intern – Advisor: Prof. Frank Rhee

- Implemented iterative versions of fuzzy clustering algorithms to improve classification accuracy and computational efficiency for large-scale datasets.

Patents

- Systems and methods for analyzing and monitoring lung function using voice and breath sound samples for respiratory care. US Patent 11,937,911 published 2024.
- Methods and systems for voice profiling as a service. US Patent App. US20210020191A1

Publications

- Amazon AGI team incl. **R Kulshreshtha**. *Amazon Nova 2: Multimodal reasoning and generation models*, 2025
- S Kapoor*, **R Kulshreshtha***. *LLM Powered After Visit Summary* (Poster, selected for Panel Discussion). American Academy of Ophthalmology, 2024. [live demo]
- MMDA Khan, PP Naval, **R Kulshreshtha**, S Venneti, A Singh. *Voice based monitoring of COPD*, CHEST Journal, 2021.
- O Ashraf, E Rabold, K Schlichtkrull, A Singh, S Venneti, MMDA Khan, **R Kulshreshtha**, PP Naval. *Voice based screening & monitoring of chronic respiratory conditions*, CHEST Journal, 2020.
- N Ryant ..., **R Kulshreshtha**. *Enhancement and Analysis of Conversational Speech: JSALT 2017*, ICASSP 2018.
- A Ravichander*, S Rijhwani*, **R Kulshreshtha***, C Nagpal, T Baltrušaitis, LP Morency. *Preserving Intermediate Objectives: One Simple Trick to Improve Learning for Hierarchical Models*, arXiv 2017.
- AA Raza, **R Kulshreshtha**, S Gella, S Blagsvedt, M Chandrasekaran, B Raj, R Rosenfeld. *Viral Spread via Entertainment and Voice-Messaging Among Telephone Users in India*, ICTD 2016.
- **R Kulshreshtha**, V Sharma, P Singh, N Agrawal, A Kumar. *Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths*, HLT-NAACL 2015.
- V Sharma, **R Kulshreshtha**, A Kumar, N Agrawal, P Singh. *Image summarization using Topic Modeling*, IEEE ICSIPA 2015.

Academic Achievements

- Represented India at the International Junior Science Olympiad 2008 (South Korea) and won a Bronze Medal.
- National Topper, Panini National Linguistics Olympiad, India 2014.
- Awarded Gandhian Young Technological Innovation Award, 2014.
- Graduate Student Fellowship at Carnegie Mellon University 2016-2017.
- Awarded Institute Merit Scholarship for topping the department, IIT Guwahati 2012.

- Won Best Project Award at the Carnegie Mellon IPTSE Winter School 2013.
- Kishore Vaigyanik Protsahan Yojana (KVPY) Scholar.
- National Talent Search Examination (NTSE) Scholar.
- Rank 24 (Team trojans), ACM-ICPC, India, 2012.

Mentorship and Community

- Reviewing - ICASSP 2026, AMLC 2023
- Invited Speaker - STEM Symposium 2020 at Fox Chapel Area High School, Pittsburgh.
- Attended Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017.
- Teaching Assistant for Deep Learning (2018) and NLP courses (2017) at Carnegie Mellon University
- Volunteer - Greater Pittsburgh Community Food Bank 2024-2025