

# Shikhar Bharadwaj

Pre-Doctoral Researcher, Google Research

 [shikhar-s.github.io](https://github.com/Shikhar-S)  [shikhar.ssu@gmail.com](mailto:shikhar.ssu@gmail.com)  Google Scholar  [github.com/Shikhar-S](https://github.com/Shikhar-S)

## EDUCATION

---

### Indian Institute of Science

M.Tech (Research) in Intelligent Systems; CGPA: 8.8/10.0

Bengaluru, India

August 2019 - May 2022

### Birla Institute of Technology and Science

B.E. (Honors) in Computer Science Engineering; CGPA: 9.74/10.0

Hyderabad, India

August 2014 - June 2018

## RESEARCH EXPERIENCE

---

### Google Research

Pre-Doctoral Researcher

Bengaluru, India

May 2022 - Present

### Microsoft Research

Research Intern

Bengaluru, India

December 2017 - May 2018

## PUBLICATIONS

S=IN SUBMISSION, C=CONFERENCE, W=WORKSHOP, \* DENOTES EQUAL CONTRIBUTION

---

### [S.1] Multimodal Modeling For Spoken Language Identification

Shikhar Bharadwaj\*, Min Ma\*, Shikhar Vashishth\*, Ankur Bapna, Sriram Ganapathy, Vera Axelrod, Siddharth Dalmia, Wei Han, Yu Zhang, Daan van Esch, Sandy Ritchie, Partha Talukdar, Jason Riesa  
Under review at ICASSP 2024

### [C.4] CodeQueries: A Dataset of Semantic Queries over Code

Surya Prakash Sahu, Madhurima Mandal, Shikhar Bharadwaj, Aditya Kanade, Petros Maniatis, Shirish Shevade  
The 17th Innovations in Software Engineering Conference (ISEC 2024)

### [C.3] Label Aware Speech Representation Learning For Language Identification

Shikhar Vashishth, Shikhar Bharadwaj, Sriram Ganapathy, Ankur Bapna, Min Ma, Wei Han, Vera Axelrod, Partha Talukdar  
INTERSPEECH 2023

### [C.2] Efficient Constituency Tree based Encoding for Natural Language to Bash Translation

Shikhar Bharadwaj and Shirish Shevade  
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)

### [C.1] Explainable Natural Language to Bash Translation using Abstract Syntax Tree

Shikhar Bharadwaj and Shirish Shevade  
Conference on Computational Natural Language Learning (CoNLL 2021)

### [W.3] MASR: Multi-Label Aware Speech Representation

Anjali Raj, Shikhar Bharadwaj, Sriram Ganapathy, Min Ma, Shikhar Vashishth  
Workshop on Automatic Speech Recognition and Understanding (ASRU 2023)

### [W.2] GitHub Issue Classification Using BERT-Style Models

Shikhar Bharadwaj\* and Tushar Kadam\*  
Second prize - Competition track at NLBSE workshop, ICSE 2022

### [W.1] An extraction based approach to keyword generation and precedence retrieval

G. V. Sandeep and Shikhar Bharadwaj  
Forum for Information Retrieval Evaluation workshop (FIRE 2017)



## SELECTED RESEARCH PROJECTS

---

### • Spoken Language Identification


Advisors: Dr. Partha Talukdar, Dr. Sriram Ganapathy, Ankur Bapna

July 2022 - September 2023

- Built **Museli [S.1]** - a multi-modal framework for language identification of YouTube videos. Our model beats the speech-only baselines by 6% (absolute F1 score) and achieves SOTA performance on public language identification datasets. This method scales even better on internal YouTube datasets spanning over 500 languages.
- The LASR model **[C.3]** uses contrastive loss in addition to MLM based losses for learning language information, resulting in an improvement by 4 F1 points.
- MASR **[W.3]**, an extension to LASR, includes external knowledge in the form of lang2vec vectors. This leads to an additional gain of 2 F1 points over LASR.
- **Natural Language to Bash Translation**  *January 2021 - March 2022*  
**Advisor:** *Dr. Shirish Shevade*
  - Developed an algorithm for translating Natural Language to Bash commands by utilizing command Abstract Syntax Tree and Bash manual page data, resulting in explainable predictions beating baselines like T5 and Seq2Seq with attention. **[C.1]**
  - Developed a novel method for Natural Language to Bash command translation using constituency tree structure of the input invocation. Results include a 1.8x improvement in inference time, 5x reduction in model parameters (compared to the Transformer) and SOTA performance. **[C.2]**
- **Project Vaani: Data collection for Indic Languages** *January 2023 - Present*  
**Advisors:** *Dr. Partha Talukdar, Dr. Sriram Ganapathy*
  - Aim: To cover the language landscape of India by region anchored speech data collection. 
  - Benchmarked and analysed results from internal ASR and Language Identification models on Vaani data.
  - Results show that Vaani is a challenging dataset for ASR models because of dialectal variations.
- **Speech to Text Transfer Learning in Multimodal models** *June 2023 - Present*  
**Advisors:** *Dr. Partha Talukdar, Dr. Sriram Ganapathy*
  - Leveraging speech for improving machine translation performance of languages with limited-text.
  - Currently evaluating on large scale **PaLM2** baselines (upto 8B parameters).
- **Semantic Queries over Code** *March 2022 - May 2022*  
**Advisor:** *Dr. Aditya Kanade, Dr. Shirish Shevade*
  - Created a benchmark for question answering over code. Implemented and evaluated models like GraphCodeBERT and CodeBERT on this dataset. **[C.4]**

## AWARDS AND RECOGNITION

---

- **Institute Merit Scholarship:** Top 2% of the batch for 7 semesters and top 1% for 1 semester at BITS
- **NTSE Scholarship:** Awarded by Govt of India for qualifying National Talent Search Examination
- **INSPIRE Scholarship:** Awarded by Govt of India to students with top 1% percentile marks
- **National Standard Examination in Physics:** Qualified NSEP (top 2.5% in India)
- **Google Internal Awards:** Four awards for contributions to the annual Google4India event, **Google-USM** and **[S.1]**
- **Media Coverage:** Audio collection effort in collaboration with the *Indian Institute of Science* has been covered by **The White House** and **The Economic Times** 


## PROFESSIONAL EXPERIENCE

---

<b>Myntra Designs Pvt Ltd.</b>	Remote, India
<i>Machine Learning Intern</i>	<i>May 2020 - July 2020</i>
<b>Media Net</b>	Mumbai, India
<i>Platform Engineer</i>	<i>July 2018 - July 2019</i>

## OTHER SKILLS

---

- **Mentorship:** TA for Computer Programming (CS F111) and Object Oriented Programming (CS F213) at BITS. Mentored a research intern at Google Research 
- **Programming Languages:** Python, C++, Java, SQL, Bash
- **Relevant Coursework:** *Graduate Level:* Deep Learning for NLP, Data Analytics, Linear Algebra and Probability, Computational Methods of Optimization, Design and Analysis of Algorithms; *Undergraduate Level:* Machine Learning, Data Mining, Data Structures and Algorithms, Object Oriented Programming
- **Tools and Frameworks:** Pytorch, Pytorch-Lightning, OpenNMT, Pandas, Matplotlib, Flax, Tensorflow