

# Sumanth Doddapaneni

## PhD in Computer Science | IIT Madras

🌐 [sumanthd17.github.io](https://sumanthd17.github.io) @ [sumanthd@cse.iitm.ac.in](mailto:sumanthd@cse.iitm.ac.in) 🐙 [github.com/sumanthd17](https://github.com/sumanthd17)  
🎓 Google Scholar 🐦 [twitter.com/sumanthd17](https://twitter.com/sumanthd17)

### Education

Jan 2022 Ongoing	<b>Indian Institute of Technology (IIT), Madras</b> Ph.D, Computer Science & Engineering Advisors: <a href="#">Mitesh M. Khapra</a> , <a href="#">Anoop Kunchukuttan</a>	Chennai, India
Aug 2016 May 2020	<b>Indian Institute on Information Technology (IIIT), Sri City</b> B.Tech., Electronics & Communications Engineering	Sri City, India

### Experience

Nov 2023 Mar 2024	<b>Google Research</b> <i>Research Intern / Hosts: <a href="#">Nitish Gupta</a>, <a href="#">Partha Talukdar</a></i> Worked on Improving Multilingual Generation in LLMs	Bangalore, India
June 2023 Oct 2023	<b>Google Research</b> <i>Student Researcher / Hosts: <a href="#">Krishna Sayana</a>, <a href="#">Vikram Aggarwal</a></i> Worked on Recommendations with Language Models	Mountain View, USA
Oct 2021 Present	<b>AI4Bharat, IIT Madras</b> <i>PhD Researcher / Advisors: <a href="#">Mitesh M. Khapra</a>, <a href="#">Anoop Kunchukuttan</a>, <a href="#">Pratyush Kumar</a></i> Working on Multilingual Language modeling and Machine Translation, with a focus on low-resource Indian languages	Chennai, India
Nov 2022 Mar 2023	<b>Mila - Quebec AI Institute</b> <i>Collaborator / Host: <a href="#">Rahul Aralikkatte</a>, <a href="#">Jackie Chi Kit Cheung</a></i> Exploring pretraining methods to develop better multilingual summarization models.	Remote
Oct 2020 Sep 2021	<b>Robert Bosch Centre for Data Science and AI, IIT Madras</b> <i>Post Baccalaureate Fellow / Advisors: <a href="#">Mitesh M. Khapra</a>, <a href="#">Anoop Kunchukuttan</a>, <a href="#">Pratyush Kumar</a></i> Built SOTA models for Machine Translation (IndicTrans) and Automatic Speech Recognition (IndicWav2Vec) for Indian Languages	Chennai, India

### Select Publications

P=Preprints, C=Conference, W=Workshop, J=Journal, \*=Equal Contribution

- [P] **Finding Blind Spots in Evaluator LLMs with Interpretable Checklists** [🔗][Code]  
[Sumanth Doddapaneni](#)\*, Mohammed Safi Ur Rahman Khan\*, Sshubam Verma, Mitesh M. Khapra  
*Under Review* [Arxiv 2024]
- [C] **IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages** [🔗][Code]  
Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, [Sumanth Doddapaneni](#), Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M. Khapra  
*62<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand* [ACL 2024]
- [J] **IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages** [🔗][Code]  
Jay Gala, Pranjal A Chitale, Raghavan AK, [Sumanth Doddapaneni](#), Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, Anoop Kunchukuttan  
*Transactions on Machine Learning Research* [TLMR 2023]
- [J] **Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages** [🔗][Code]  
Gowtham Ramesh\*, [Sumanth Doddapaneni](#)\*, et. al  
*Transactions of the Association for Computational Linguistics* [TACL 2022]

- [J] **A Survey in Adversarial Defences and Robustness in NLP** [🔗]  
Shreya Goyal, [Sumanth Doddapaneni](#), Mitesh M Khapra, Balaraman Ravindran  
*ACM Computing Surveys* [CSUR 2023]
- [C] **Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages** [🔗][Code]  
[Sumanth Doddapaneni](#), Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar  
*61<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, Toronto, Canada* [ACL 2023]
- [C] **Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages** [🔗]  
Arnav Mhaske, Harshit Kedia, [Sumanth Doddapaneni](#), Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy V, Anoop Kunchukuttan  
*61<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, Toronto, Canada* [ACL 2023]
- [C] **Vārta: A Large-Scale Headline-Generation Dataset for Indic Languages** [🔗]  
Rahul Aralikkatte, Ziling Cheng, [Sumanth Doddapaneni](#), Jackie Chi Kit Cheung  
*Findings of 61<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, Toronto, Canada* [ACL 2023]
- [C] **Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages** [🔗]  
Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, [Sumanth Doddapaneni](#), Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra  
*48<sup>th</sup> IEEE International Conference on Acoustics, Speech, and Signal Processing, Rhodes Island, Greece* [ICASSP 2023]
- [C] **Towards Building ASR Systems For The Next Billion Users** [🔗][Code]  
Tahir Javed, [Sumanth Doddapaneni](#), Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra  
*36<sup>th</sup> AAAI Conference on Artificial Intelligence, Vancouver, Canada* [AAAI 2022]
- [P] **A Primer on Pretrained Multilingual Language Models** [🔗]  
[Sumanth Doddapaneni\\*](#), Gowtham Ramesh\*, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar  
*Under Review* [ArXiv]

## Honours and Grants

---

**Google** PhD Fellowship 2023 ([List of Recipients](#))

**Google Research** Selected to attend the Google Research Week 2022, 2023

**Naver Labs and Univ. Grenoble Alpes** Selected to attend the ALPS Winter School 2022

**TFRC Grant** Received TPU Research credits to carry out work on LMs 2022

**MSR Travel Grant** Received Microsoft Research Travel grant to attend ACL 2022

**Robert Bosch Centre** Received the Post Baccalaureate Fellowship to work on interdisciplinary AI 2021

## Select Research Projects

---

**Multilingual Summarisation** [🔗][Code] Nov'22 - Mar'23

*Advisors: Dr. Rahul Aralikkatte, Dr. Jackie Chi Kit Cheung*

- > Studying the effect of pretraining in multilingual summarisation
- > Pretraining and fine-tuning generative models for abstractive summarisation
- > Accepted to Findings of ACL 2023

**Multilingual Language Modeling** [🔗][Code] Oct'21 - Jan'23

*Advisors: Dr. Mitesh M. Khapra, Dr. Anoop Kunchukuttan, Dr. Pratyush Kumar*

- > Building Large scale monolingual corpora and Language Models for Indian Languages
- > Building strong testsets to evaluate the zero-shot generalisation of the models
- > Studying the various objectives and data regimes that improve zero-shot generalisation of the models
- > Understanding the efficacy of adapters in zero-shot generalisation of the LMs
- > Work published at ACL 2023

## Speech Recognition [Try the Model][Code]

Jun'21 - Feb'22

Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#)

- > Building state-of-the-art models for Automatic Speech Recognition for Indian languages
- > Understanding the various effects of LM on downstream ASR task
- > Accepted as a long paper at AAAI 2022

## Machine Translation [Try the model][Code]

Feb'21 - Oct'21

Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#)

- > Built state of the art translation model for Indian languages to English and vice versa
- > Created the largest bilingual corpora (~50M pairs) across 11 languages for NMT training.
- > Work published in TACL (Volume 10, 2022)

## Academic Service

---





<b>Reviewer</b>	NAACL'24, EMNLP'23 (Best Reviewer), ACL'23; ACL Rolling Review 2022, 2023, 2024
<b>Volunteer</b>	EACL'21, ICML'21, NeurIPS'21, EMNLP'21, ACL'22, EMNLP'22

## Talks & Volunteering Roles

---

<b>Invited Talk, Google Research India</b>	<i>Speaker</i>	Feb 2023
> Talk on "Building Natural Language Understanding (NLU) capabilities for Indic languages"		
<b>Invited Talk, Swiggy Data Science</b>	<i>Speaker</i>	Jan 2022
> Talk on "Towards Building ASR Systems for the Next Billion Users"		
<b>Organizer at NLP with Friends</b>	<i>Organizer</i>	Jan 2023 - Present
<b>NLP Reading Group, AI4Bharat</b>	<i>Organizer</i>	Jan 2022 - Dec 2022

## References

- 
- > Dr. Mitesh M. Khapra ..... Associate Professor, IIT Madras, India 
  - > Dr. Anoop Kunchukuttan ..... Principal Applied Researcher, Microsoft, India 
  - > Dr. Pratyush Kumar ..... Co-Founder, Sarvam, India 
  - > Dr. Rahul Aralikkatte ..... Sarvam, India 
  - > Dr. Raj Dabre ..... Researcher, NICT, Japan 