

Sumanth Doddapaneni

ML Researcher | Sarvam AI

🌐 sumanthd17.github.io @ sumanth@sarvam.ai 🌐 github.com/sumanthd17
🎓 Google Scholar 🐦 twitter.com/sumanthd17

Education

On Leave Jan 2022	Indian Institute of Technology (IIT), Madras Ph.D, Computer Science & Engineering Advisors: Mitesh M. Khapra , Anoop Kunchukuttan	Chennai, India
May 2020 Aug 2016	Indian Institute on Information Technology (IIIT), Sri City B.Tech., Electronics & Communications Engineering	Sri City, India

Experience

Present Apr 2025	Sarvam AI ML Researcher Building Multilingual LLMs.	Bangalore, India
Mar 2024 Nov 2023	Google Research Research Intern / Hosts: Nitish Gupta , Partha Talukdar Worked on Improving Multilingual Generation in LLMs	Bangalore, India
Oct 2023 June	Google Research Student Researcher / Hosts: Krishna Sayana , Vikram Aggarwal Worked on Recommendations with Language Models	Mountain View, USA
Apr 2025 Oct 2021	AI4Bharat, IIT Madras PhD Researcher / Advisors: Mitesh M. Khapra , Anoop Kunchukuttan , Pratyush Kumar Working on Multilingual Language modeling and Machine Translation, with a focus on low-resource Indian languages	Chennai, India
Sep 2021 Oct 2020	Robert Bosch Centre for Data Science and AI, IIT Madras Post Baccalaureate Fellow / Advisors: Mitesh M. Khapra , Anoop Kunchukuttan , Pratyush Kumar Built SOTA models for Machine Translation (IndicTrans) and Automatic Speech Recognition (IndicWav2Vec) for Indian Languages	Chennai, India

Select Publications



Full List

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

- [C] **Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs** [🔗][Code][Models]
[Sumanth Doddapaneni](#)*, [Mohammed Safi Ur Rahman Khan](#)*, [Dilip Venkatesh](#), [Raj Dabre](#), [Anoop Kunchukuttan](#),
[Mitesh M. Khapra](#)
The 63rd Annual Meeting of the Association for Computational Linguistics [ACL 2025]
- [C] **[Outstanding Paper] Finding Blind Spots in Evaluator LLMs with Interpretable Checklists** [🔗][Code][Data]
[Sumanth Doddapaneni](#)*, [Mohammed Safi Ur Rahman Khan](#)*, [Sshubam Verma](#), [Mitesh M. Khapra](#)
The 2024 Conference on Empirical Methods in Natural Language Processing [EMNLP 2024]
- [C] **[Outstanding Paper] 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](#)
62nd 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 [TMLR 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]
- [C] **Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages** [🔗][Code]
[Sumanth Doddapaneni](#), Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar
61th Annual Meeting of the Association for Computational Linguistics, Toronto, Canada [ACL 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
36th AAAI Conference on Artificial Intelligence, Vancouver, Canada [AAAI 2022]

Honours and Grants

Outstanding Paper Award @ EMNLP 2024 Finding Blind Spots in Evaluator LLMs with Interpretable Checklists

Outstanding Paper Award @ ACL 2024 IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages

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

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

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

Auto Evaluation with LLMs [Code][Models] Apr'24 - Ongoing
Advisors: [Dr. Mitesh M. Khapra](#)

- > Analyzed failure modes in **automated evaluators** using systematic perturbations, identifying significant shortcomings in existing evaluation methods.
- > Developed the **first cross-lingual auto-evaluator** and released a **multilingual benchmark**, enabling comprehensive evaluation of multilingual LLMs.
- > Published results at **EMNLP 2024** [[Outstanding Paper Award](#)].

Machine Translation [Try the model][Code] Feb'21 - May'23
Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#)

- > Developed state-of-the-art translation models for Indian languages, releasing two versions—**IndicTrans v1** and **v2**—over multiple years to enhance multilingual performance.
- > Built the largest bilingual corpus, **Samanantar** (~ 50M sentence pairs), spanning 11 Indian languages for NMT training. This effort was further expanded in the v2 release with **BPCC** (~ 230M sentence pairs) across 22 languages.
- > Published papers at **TACL** and **TMLR**.

Language Modeling [🔗] Dec'21 - Feb'24
Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#)

- > Developed the largest monolingual corpus, **IndicCorp v2**, followed by an even larger version, **Sangraha**, in the subsequent release.
- > Built language models, including **IndicBERT v2** for masked language modeling and **VartaT5** for text generation, and created comprehensive multilingual benchmarks for Indian languages, **IndicXTREME**.
- > Published multiple papers at **ACL 2023** and **ACL 2024**, with IndicLLMSuite receiving the [[Outstanding Paper Award](#)].

Academic Service

Reviewer NAACL: 2024, 2025, EMNLP: 2023 (Best Reviewer), 2024, ACL: 2023, 2024;

Volunteer EACL'21, ICML'21, NeurIPS'21, EMNLP'21, ACL'22, EMNLP'22