Anil K. Ramakrishna

anil.k.ramakrishna@gmail.com — anilkramakrishna.github.io

Research Interests

Responsible AI, Large Language Models, Natural Language Processing, Machine Learning.

Experience

Research Scientist August 25 - present
Meta Superintelligence Labs (MSL).
Menlo Park, CA

- I work in the Trust and Safety team of MSL.

Senior Applied Scientist July 23 - August 25
AGI Foundations, Amazon Inc.
Los Angeles, CA

- I worked in the Responsible AI team in the AGI Foundations organization where my research focused on LLM unlearning, high quality data curation, LLM model alignment and factuality, and robust model evaluations.

Applied Scientist October 19 - July 23
Alexa AI. Amazon Inc.
Los Angeles, CA

- Worked on Data Efficient Learning (for privacy), Interpretable Model Building and Defect Reduction within the Natural Language Understanding subsystem of Alexa Virtual Assistant.

Applied Scientist Intern May 18 - August 18
Amazon Inc.
Boston, MA

- Developed a model combination strategy to address privacy constraints with shared data training for Amazon's Alexa NLU pipeline.

Education

Doctor of Philosophy May 14 - October 19
Computer Science — *GPA: 3.8/4.0*
Advisor: Prof. Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles, CA

Master of Science
Electrical Engineering — *GPA: 3.95/4.0* Fall 18 - Spring 19
Computer Science — *GPA: 3.7/4.0* Fall 12 - Spring 14
University of Southern California, Los Angeles, CA

Bachelor of Engineering July 06 - May 10
Information Science — *Agg. %: 80.2/100*
B.M.S. College of Engineering
Visvesvaraya Technological University, Belgaum, India

Patents

Anil K Ramakrishna, Rahul Gupta, Yuval Merhav, Zefei Li, Heather Brooke Spetalnick. *Machine learning model updating.* US Patent #11978438, 2024/5/7.

Shrikanth Narayanan, Victor Martinez Palacios, **Anil Ramakrishna**, Krishna Somandepalli, Nikolaos Malandrakis, Karan Singla, *Linguistic analysis of differences in* *portrayal of movie characters.* US Patent #11775765, 2023/10/3.

Anoop Kumar, **Anil K Ramakrishna**, Sriram Venkatapathy, Rahul Gupta, Sankaranarayanan Ananthakrishnan, Premkumar Natarajan. *Spoken language understanding models.* US Patent #11574637, 2023/2/7.

Shrikanth Narayanan, Victor Martinez Palacios, **Anil Ramakrishna**, Krishna Somandepalli, Nikolaos Malandrakis, Karan Singla, *Linguistic analysis of differences in portrayal of movie characters.* US Patent #10956679, 2021/3/23.

An additional eight patents are filed and are currently being reviewed by the US patent office.

Academic Service

Organizing Committee
Ethics Chair at EMNLP 2025.
Workshop on Trustworthy NLP, NAACL 2025.
SemEval 2024 Task 4: Unlearning sensitive content from Large Language Models.
Satellite Workshop on Trustworthy Speech Processing, ICASSP 2024.
Special Session on Trustworthy Speech Processing, Interspeech 2022.
Workshop on Trustworthy NLP, NAACL 2021.

Program Committee
2025: ARR (February), ICCV, SemEval 2026
2024: COLM, ARR (April, June, August, October, December).
2023: ACL, ICASSP.
2022: EMNLP, ARR (June), ICASSP, AAAI.
2021: NAACL, EMNLP, ACL-IJCNLP, EACL.
2020: Privacy Preserving ML Workshop (Amazon Machine Learning Conference), EMNLP, The Ninth Joint Conference on Lexical and Computational Semantics, COLING Industrial Track, Palgrave Communications, AAAI.
2019: PLOS One, NAACL.
2018: IEEE Transactions on Affective Computing.

Recent Publications

Taha Entesari, Arman Hatami, Rinat Khaziev, **Anil Ramakrishna**, Mahyar Fazlyab. Constrained Entropic Unlearning: A Primal-Dual Framework for Large Language Models. NeurIPS 2025.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, Rahul Gupta. LUME: LLM Unlearning with Multitask Evaluations. EMNLP 2025.

Yixin Wan, **Anil Ramakrishna**, Kai-Wei Chang, Volkan Cevher, Rahul Gupta. Not Every Token Needs Forgetting: Selective Unlearning for Balancing Forgetting and Utility in Large Language Models. EMNLP 2025.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, Rahul Gupta. SemEval-2025 Task 4: Unlearning sensitive content from Large Language Models. SemEval 2025.

Chongyu Fan, Jinghan Jia, Yihua Zhang, **Anil Ramakrishna**, Mingyi Hong, Sijia Liu. Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond. ICML 2025.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, **Anil Ramakrishna**, Tagyoung Chung. *REAL Sampling: Boosting Factuality and Diversity of Open-Ended Generation via Asymptotic Entropy.* Transactions of the Association for Computational Linguistics, 2025.

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, **Anil Ramakrishna**, Kai-Wei Chang, Volkan Cevher, Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. NAACL, 2025.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, **Anil Ramakrishna**, Dimitrios Dimitriadis, Jieyu Zhao, Salman Avestimehr. Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs. NAACL, 2025.

Anubrata Das, Manoj Kumar, Ninareh Mehrabi, **Anil Ramakrishna**, Anna Rumshisky, Kai-Wei Chang, Aram Galstyan, Morteza Ziyadi, Rahul Gupta. On Localizing and Deleting Toxic Memories in Large Language Models. NAACL, 2025.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, **Anil Ramakrishna**, Tagyoung Chung. *by Extrapolating the Probabilities of a Huge and Hypothetical LM.* EMNLP, 2024.

Tao Meng, Ninareh Mehrabi, Palash Goyal, **Anil Ramakrishna**, Aram Galstyan, Richard Zemel, Kai-Wei Chang, Rahul Gupta, Charith Peris. *Attribute Controlled Fine-tuning for Large Language Models: A Case Study on Detoxification.* EMNLP, 2024.

Elan Markowitz, **Anil Ramakrishna**, Jwala Dhamala, Ninareh Mehrabi, Charith Peris, Rahul Gupta, Kai-Wei Chang, Aram Galstyan. *Tree-of-Traversals: A Zero-Shot Reasoning Algorithm for Augmenting Black-box Language Models with Knowledge Graphs.* ACL, 2024.

Anil Ramakrishna, Rahul Gupta, Jens Lehmann, Morteza Ziyadi. *Invite: a testbed of automatically generated invalid questions to evaluate large language models for hallucinations.* EMNLP, 2023.

Rahul Sharma*, **Anil Ramakrishna***, Ansel MacLaughlin, Anna Rumshisky, Jimit Majmudar, Clement Chung, Salman Avestimehr, Rahul Gupta. *Federated learning with noisy user feedback.* NAACL, 2021.

Anil Ramakrishna, Rahul Gupta, Shrikanth Narayanan, *Joint Multi-Dimensional Model for Global and Time-Series Annotations*, in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.3006418.

A complete list of publications can be found in my Google Scholar page.

Computing Skills

Languages:	Python, BASH shell scripting, C, C++ & Java(Beginner), LaTeX.
Technical Computing:	MATLAB, GNU Octave, R, LabVIEW(Beginner).
System Administration:	GNU/Linux, openSUSE.

Service

I'm a certified yoga (200Hour) and breath-work based meditation (1000Hour) instructor. I spend my free time facilitating wellness sessions in university campuses.

Certified Facilitator, Art of Living Foundation	Jan 19 - present
Coordinator, Machine Learning Reading Group, SAIL	Fall 15 - Summer 19
President, Yoga and Meditation Club at USC	Fall 14 - Summer 19
Vice-President, Yoga and Meditation Club at USC	Fall 13 - Spring 14
Volunteer, Art of Living Foundation	07 - present

Nationality Indian (with US Permanent Residency)