Nishant Subramani

nishant2@andrew.cmu.edu https://nishantsubramani.github.io

Biography and Research Interests

I'm a PhD student at CMU in the Language Technologies Institute advised by Mona Diab. I'm interested in model interpretability and understanding, more specifically understanding the internals of language models with an eye towards controllability: steering their generations in a reliable, trustworthy, and efficient manner. I wrote the first papers on steering vectors both in LSTM and transformer-based language models. I also have experience working on LLM initiatives such as BLOOM and OLMo, winning two best paper awards at ACL 2024. I have worked on numerous topics in NLP, ML, Speech, Computer Vision, and Causality leading to over 20 publications with over 4000 total citations at conferences including NeurIPS, ICML, COLM, AAAI, FAccT, ACL, NAACL EMNLP, and TACL.

Education

- **Carnegie Mellon University | Language Technologies Institute (LTI)** Aug 2023 – Present
  - **PhD Language Technologies (Computer Science)** GPA: 4.1/4.0
  - **PhD Advisor:** Mona Diab
  - **Graduate Courses**: Neural Code Generation, Talking to Robots, Subword Modeling, Multimodal Machine Learning
- **Courant Institute of Mathematical Sciences | New York University** Sept 2017 – May 2019
  - **M.S. Computer Science** (Deep Learning & NLP) GPA: 3.8/4.0
  - **Research Advisors:** Kyunghyun Cho and Sam Bowman
  - **Graduate Courses**: Deep Learning, Deep Generative Models, Deep Learning for NLP
- **Northwestern University** Sept 2013 – June 2017
  - **B.A./M.S. Statistics/Computer Science**; Stat GPA: 4.0/4.0; MS GPA: 4.0/4.0
  - **Research Advisor:** Doug Downey
  - **Graduate Courses**: Deep Learning, Machine Learning Foundations, Probabilistic Graphical Models, Data Mining, Adv Topics in ML, Statistical Pattern Recognition, Computational Learning Theory, Adv Topics in Bayesian Stats

Work Experience

- **Student Researcher | Google (Cloud AI Research)** May 2025 – Present
  - Advised by Hamid Palangi on actionable mechanistic interpretability.
- **Research Scientist Intern | Microsoft Research (Semantic Machines)** June 2024 – August 2024
  - Advised by Sam Thomson and Yu Su.
  - Developed MICE for CATs: Model-Internal Confidence Estimation for Calibrating Agents with Tools (NAACL2025).
- **Predoctoral Young Investigator | Allen Institute for AI** June 2021 – October 2023
  - Advised by Matthew Peters.
  - Developed a method to steer pretrained language models using fixed-length steering vectors - ACL2022 Findings.
  - Worked on data governance as part of the BigScience project leading to a paper published at FAccT 2022.
  - Working on efficient controllable text generation (in prep), data-driven approaches for benchmark design and dataset complexity measurement (in prep), and personal information auditing in large web corpora (TrustNLP workshop at ACL23).
  - Co-hosted the NLP Highlights Podcast with Alexis Ross on episodes related to PhD applications in NLP.
- **Predoctoral Resident | Intel Intelligent Systems Lab** January 2021 – June 2021
  - Advised by Vladlen Koltun.
  - Developed methods for the analysis and controllability of large pretrained language models.
- **Machine Learning Research Scientist | Scale AI** April 2020 – December 2020
  - Led ML research for NLP and developed multi-task optical character recognition & document understanding models.
  - Published a survey paper on document understanding at the MLRSA workshop at NeurIPS 2020.
  - Created a natural adversarial objects dataset to improve robustness of object detectors, paper accepted to the Data Centric AI workshop at NeurIPS 2021.
- **Research Scientist | AI Foundation** July 2019 – January 2020
  - Built a sample- and memory-efficient multi-task fake speech detection system and published at AAAI 2020.
  - Created a large, diverse fake speech dataset to improve internal fake speech detection systems.
  - Developed an audio-driven facial animation model, which made AI rendered puppets more realistic.

Selected Publications (with embedded links)

25. **Model Internal Sleuthing: Finding Lexical Identity and Inflectional Morphology in Modern Language Models**
    *Michael Li, **Nishant Subramani**.*
    COLM 2025 (Interplay Workshop)
24. **LLM Microscope: What Model Internals Reveal About Answer Correctness and Context Use**
    Jiarui Liu, Jivitesh Jain*, Mona Diab, **Nishant Subramani**.*
    COLM 2025 (Interplay Workshop)

23. **SimBA: Simplifying Benchmark Analysis Using Performance Matrices Alone**
    Nishant Subramani*, Alfredo Gomez*, Mona Diab.
    ICML 2025 (R2FM Workshop)

22. **Personal Information Parroting in Language Models**
    Nishant Subramani, Kshitish Ghate, Mona Diab.
    ICML 2025 (MEMFM Workshop); ACL 2025 (L2M2 Workshop)

21. **MICE for CATs: Model-Internal Confidence Estimation for Calibrating Agents with Tools**
    Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, Sam Thomson.
    NAACL2025; Best Paper Runner-Up at CMU Student Research Symposium

20. **OLMo: Accelerating the Science of Language Models**
    Dirk Groeneveld, Iz Beltagy, ... ***Nishant Subramani***, ... Luca Soldaini, Noah A. Smith, Hannaneh Hajishirzi.
    Best Paper Award at ACL2024

19. **Dolma: An Open Corpus of 3 Trillion Tokens for Language Model Pretraining Research**
    Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, ... ***Nishant Subramani***, ... Kyle Lo.
    Best Paper Award at ACL2024

18. **Evaluating Personal Information Parroting in Language Models**
    Nishant Subramani*, Kshitish Ghate*, Mona Diab.
    NAACL2024 (TrustNLP Workshop)

17. **Robust Tooling and New Resources for Large Language Model Evaluation via Catwalk**
    Kyle Richardson, Ian Magnusson, Oyvind Tafjord, ... ***Nishant Subramani.***
    EMNLP 2023 (GEM Workshop; Extended Abstract)

16. **Detecting Personal Information in Training Corpora: an Analysis**
    Nishant Subramani*, Alexandra Sasha Luccioni*, Jesse Dodge, and Margaret Mitchell.
    ACL 2023 (TrustNLP Workshop)

15. **Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search**
    Daniel King*, Zejiang Shen*, ***Nishant Subramani***, Daniel S. Weld, Iz Beltagy, and Doug Downey.
    EMNLP 2022 (GEM Workshop)

14. **GEMv2: Multilingual NLG Benchmarking in a Single Line of Code**
    Sebastian Gehrmann, ... ***Nishant Subramani***, ... Yufang Hou.
    EMNLP 2022

13. **Extracting Latent Steering Vectors from Pretrained Language Models**
    Nishant Subramani, Nivedita Suresh, and Matthew E. Peters.
    ACL 2022 Findings

12. **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model**
    Teven Le Scao, ... ***Nishant Subramani***, ... Thomas Wolf.
    BigScience Workshop

11. **Data Governance in the Age of Large-Scale Data-Driven Language Technology**
    Yacine Jernite, Huu Nguyen, ... ***Nishant Subramani***, ... Margaret Mitchell.
    FAccT 2022

10. **Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets**
    Julia Kreutzer, Isaac Caswell, ... ***Nishant Subramani***, ... Mofetoluwa Adeyemi.
    TACL 2022

9. **The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics**
    Sebastian Gehrmann, ... ***Nishant Subramani***, ... Jiawei Zhou.
    ACL 2021 (GEM Workshop)

8. **Natural Adversarial Objects**
    Felix Lau, ***Nishant Subramani***, Alexandra Harrison, Aerin Kim, Elliot Branson, and Rosanne Liu.
    NeurIPS 2021 (Data Centric AI Workshop)

7. **Discovering Useful Sentence Representations from Large Pretrained Language Models**
    Nishant Subramani and Nivedita Suresh.
    arXiv 2020

6. **A Survey of Deep Learning Approaches for OCR and Document Understanding**
    Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam.
    NeurIPS 2020 (MLRSA Workshop)

5. **Learning Efficient Representations for Fake Speech Detection**
    Nishant Subramani and Delip Rao.
    AAAI 2020

4. **Can Unconditional Language Models Recover Arbitrary Sentences?**
   Nishant Subramani, Samuel R. Bowman, and Kyunghyun Cho.
   NeurIPS 2019

3. **Pag2admg: An Algorithm for the Complete Causal Enumeration of a Markov Equivalence Class**
   Nishant Subramani
   ICML 2018 (CausalML Workshop)

2. **PAG2ADMG: A Novel Methodology to Enumerate Causal Graph Structures**
   Nishant Subramani, and Doug Downey.
   AAAI 2017 (Student Abstract)

1. **Identifying the Best Predictors of Unmet Health Care Needs in Children with DBD.**
   Nishant Subramani
   Northwestern Undergraduate Research Journal 2015.

Research Experience

- **Researcher | The Big Science Initiative** July 2021 – May 2022
  - Worked with the Data Governance group to publish a FAccT 2022 paper on data governance pertaining how to govern the data & models that were collected, created, or utilized for building BLOOM, a 176B open-access multilingual language model.
- **Research Collaborator | Allen Institute for AI** October 2020 – March 2021
  - Worked with Doug Downey and Daniel King on scientific concept generation models from scientific papers with the Semantic Scholar team.
- **NLP Researcher | Masakhane** May 2020 – May 2022
  - Co-organized the AfricaNLP 2021 workshop. **Accepted at EACL2021**
- **Research Assistant | New York University** September 2017 – May 2019
  - Advised by Kyunghyun Cho and Sam Bowman.
  - Developed a framework to analyze the sentence space of a recurrent neural language model.
  - Built a pipeline to investigate using a language model as a universal decoder for multitask natural language generation.
- **Deep Learning Research Intern | Salesforce Research** March 2017 – August 2017
  - Supervised by Richard Socher
  - Built a multitask NLP system trained end-to-end for a vareity of NLP tasks.
  - Investigated impact of CoVe pretraining on state of the art abstractive summarization and question answering models.
- **Research Assistant | Northwestern University** July 2014 – March 2015; March 2016 – June 2017
  - Advised by Doug Downey.
  - Improved my *pag2admg* algorithm developed at ETH Zurich into a method that generates all Markov equivalent acyclic directed mixed graphs (not necessary just ancestral) from a PAG.
  - Developed various methodologies to identify deep net hyperparameter settings more efficiently using active learning and sampling.
  - Developed various ensembling methodologies to improve state-of-the-art language model performance on the Penn Tree Bank dataset.
  - Developed alternative dropout methodologies to increase variance of models from epoch to epoch to improve deep neural network performance on a variety of tasks.
  - Developed methods to input pre-existing analogical knowledge to improve word-embeddings in Google's word2vec models.
  - Developed methods to utilize importance sampling to help stochastic gradient descent convergence for neural sentence-level language modeling.
- **Research Assistant in Biomedical Informatics | Stanford University** Jun 2015 – Jan 2016
  - Supervised by Olivier Gevaert
  - Developed a Bayesian Network structure learning methodology to identify a genetic basis for Glioblastoma.
- **Research Assistant in Biomedical Informatics | Feinberg School of Medicine** Jan 2016 – March 2016
  - Supervised by Yuan Luo
  - Predicted ICU 30-day readmission rates from a multivariate panel of physiological measurements using Subgraph Augmented Non-Negative Matrix Factorization (SANMF).
- **Master's Semester Project Student in Systems Biology | ETH Zurich** Sept 2015 – Jan 2016
  - Supervised by Manfred Claassen
  - Developed a methodology (*The Boundary Searcher*) to efficiently calculate the r-convex hull of a point cloud in high dimensions.
- **Master's Semester Project Student in Statistics | ETH Zurich** Sept 2015 – Jan 2016
  - Supervised by Marloes Maathuis
  - Developed a novel methodology to transform a given partial ancestral graph (PAG) to the set of all ancestral acyclic directed mixed graphs that belong in the Markov equivalence class that the PAG encodes.

Teaching Experience

- **Teaching Assistant for Advanced Natural Language Processing | CMU** Jan 2024 – May 2024

  - Graduate Course: CS 11-711 - Advanced Natural Language Processing.
  - Give a lecture on mechanistic interpretability and steering vectors.
  - Developed homework assignments and running recitations on modern NLP software packages such as vLLM.
  - Helped advise research projects completed by students in the course that involve modern NLP.
- **Teaching Assistant for Natural Language Understanding | NYU** Jan 2018 – May 2018
  - Graduate Course: DSGA-1012 - Natural Language Understanding.
  - Gave a lecture on deep learning fundamentals for NLU.
  - Developed homework assignments and ran the tutorial sessions of the course.
  - Helped advise research projects completed by students in the course that involved deep learning applied to language.
- **Teaching Assistant for Statistical Language Modeling | Northwestern** Jan 2017 – Mar 2017
  - Graduate Course: EECS 496 - Statistical Language Modeling focusing on Deep Learning.
  - Constructed seminar reading list; helped other students understand seminal deep NLP papers.
- **Teaching Assistant for Probabilistic Graphical Models | Northwestern** Sept 2016 – Dec 2016
  - Graduate Course: EECS 495 - Probabilistic Graphical Models.
  - Helped to design course materials and structure for this graduate course.
  - Developed and graded assignments; held office hours.
- **Teaching Assistant for Mathematical Foundations of CS | Northwestern** Sept 2016 – Dec 2016
  - Undergraduate Course: EECS 212 - Mathematical Foundations of Computer Science.
  - Helped to develop and grade assignments and exams; held office hours.
- **Teaching Assistant for Machine Learning | Northwestern** Feb 2016 – June 2016
  - Undergraduate/Graduate Course: EECS 349 - Machine Learning.
  - Devised methodology for and built a mechanical TA which uses the Vancouver crowd sourcing algorithm.
  - Helped to design tree search and decision tree assignments, graded assignments, and held office hours.
- **Co-Instructor for Computing Applications I & II | Northwestern** Sept 2014 – March 2015
  - Undergraduate Courses: ISP 101-1, 101-2 - Computing Applications I/II.
  - Co-taught course with three other teaching assistants.
  - Wrote exam questions and assignments covering python and R basics.

Mentorship Experience

- **Students Mentored**
  - Michael Li (Undergraduate Student at CMU) April 2025 – Present
    - COLM 2025 workshop paper.
  - Jiarui Liu (MIIS -> MLT Student at CMU) -> PhD Student CMU LTI March 2024 – Present
    - COLM 2025 workshop paper.
  - Kshitish Ghate (MIIS -> MLT Student at CMU) -> PhD Student UnivWashington CSE August 2023 – December 2024
    - ACL 2025 workshop paper.
    - ICML 2025 workshop paper.
    - TrustNLP 2024 workshop paper.

Other Experience

- **Deep Learning Consultant | Talkspace** November 2017 – August 2018
  - Taught Talkspace's Data Science team about deep learning fundamentals and helped build domain-specific models.

Invited Talks

- ***MICE for CATs: Model-Internal Confidence Estimation for Calibrating Agents with Tools*** April 2025
  CMU LTI Student Research Symposium. Pittsburgh, PA.
- ***Steering Vectors: an alternative way to steer language models*** November 2023
  Ontario Tech University. Virtual.
- ***Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*** May 2022
  ACL 2022, Dublin, Ireland.
- ***Fantastic Continuous-valued Sentence Representations and How to Find Them.*** July 2021
  ISI NLP Seminar. Virtual.
- ***A Survey of Deep Learning Approaches for OCR and Document Understanding.*** December 2020
  NeurIPS MLRSA Workshop 2020, Virtual.
- ***Can Unconditional Language Models Recover Arbitrary Sentences?*** March 2020
  SRI International, Menlo Park, CA.
- ***PAG2ADMG***. February 2017
  AAAI 2017, San Francisco, CA. Student Abstract Spotlight.
- ***How Evil are Turnovers?*** June 2014
  Undergraduate Research Expo, Northwestern University.

Invited Posters

- ***Personal Information Parroting in Language Models*** July 2025
  ACL 2025 (L2M2 Workshop), Vienna, Austria. Poster.
- ***Personal Information Parroting in Language Models*** July 2025
  ICML 2025 (MEMFM Workshop), Vancouver, Canada. Poster.
- ***SimBA: Simplifying Benchmark Analysis Using Performance Matrices Alone*** July 2025
  ICML 2025 (R2FM Workshop), Vancouver, Canada. Poster.
- ***MICE for CATs: Model-Internal Confidence Estimation for Calibrating Agents with Tools*** April 2025
  NAACL 2025, Albuquerque, USA. Poster.
- ***Evaluating Personal Information Parroting in Language Models*** June 2024
  NAACL 2024 (TrustNLP Workshop), Mexico City, Mexico. Poster.
- ***Detecting Personal Information in Training Corpora: an Analysis*** July 2023
  ACL 2023 (TrustNLP Workshop), Toronto, Canada. Poster.
- ***Don't Say What You Don't Know*** December 2022
  EMNLP 2022 (GEM Workshop), Abu Dhabi, UAE. Poster.
- ***Extracting Latent Steering Vectors from Pretrained Language Models*** May 2022
  ACL 2022 Findings, Dublin, Ireland. Poster.
- ***Learning Efficient Representations for Fake Speech Detection.*** February 2020
  AAAI 2020, New York, USA. Poster.
- ***Can Unconditional Language Models Recover Arbitrary Sentences?*** December 2019
  NeurIPS 2019, Vancouver, Canada. Poster.
- ***PAG2ADMG***. ICML 2018, Stockholm, Sweden. Causal ML Workshop. Poster. July 2018
- ***PAG2ADMG***. AAAI 2017, San Francisco, CA. Student Abstract Poster. February 2017
- ***Pag2Admg***. Undergraduate Research Expo, Northwestern University. Poster. June 2016
- ***The Boundary Searcher***. EECS Poster Fair, Northwestern University. Poster. Apr 2016
- ***Predicting Unmet Health Care Needs in Children with DBD.*** June 2015
  Undergraduate Research Expo, Northwestern University. Poster.
- ***Predicting Unmet Health Care Needs in Children with DBD.*** Mar 2015
  EECS Poster Fair, Northwestern University. Poster.
- ***How Evil are Turnovers?*** Apr 2014
  Computational Statistics Conference, Northwestern University. Poster.

Professional Service

- Workshop Organizer for AfricaNLP at **EACL 2021** 2021
- Conference Reviewer for ACL Rolling Review 2021 – Present
- Conference Reviewer for COLM 2021 – Present
- Conference Reviewer for NeurIPS 2017 – Present
- Conference Reviewer for ICLR 2019 – Present
- Conference Reviewer for ICML 2020 – Present
- Conference Reviewer for ACL (moved to ARR) 2021 – 2023
- Conference Reviewer for EMNLP (moved to ARR) 2019 – 2023
- Conference Reviewer for AAAI 2020 – 2021
- Conference Reviewer for CVPR 2021
- Conference Reviewer for ICCV 2017
- Workshop Reviewer for ICML MEMFM Workshop 2025
- Senior Area Chair for SustaiNLP Workshop at **EMNLP 2022** 2022
- Program Committee Member for GEM Workshop at **ACL 2021** 2021

Other Service

- CMU Language Technologies Institute Mentorship Program Mentor for PhD students in LTI 2024 – Present
- CMU Mentor for Undergraduates in Computer Science 2023 – Present
- CMU Graduate Application Support Program Mentor for Computer Science 2023 – Present

Awards & Honors

- Best Paper Runner-Up Award at CMU LTI SRS 2025 for **MICE for CATs** April 2025
- Best Paper Award at ACL 2024 for **OLMo: Accelerating the Science of Language Models** August 2024
- Best Paper Award at ACL 2024 for **Dolma: An Open Corpus of 3 Trillion Tokens** August 2024
- Waitlisted for the NDSEG Fellowship April 2024
- Henry M. MacCracken Graduate Fellowship (5 year fully-funded PhD Fellowship) September 2017 - May 2019

- Charles A & Ruby E Howell Endowed Scholarship (∼ $30,000 yearly) December 2014 - June 2017
- Inaugural ETH Zurich Exchange Program Acceptee (1 of 3 students accepted) September 2015 - February 2016
- Integrated Science Program admission from Northwestern (1 of ∼ 30 admitted across Northwestern) March 2013
- University Scholars Nomination from Penn (Undergraduate Research Program) March 2013
- Likely Letter from Cornell University (1 of ∼ 25 students across the country to receive this) March 2013
- Intel Science Talent Search (ISTS) Outstanding Written Report Award March 2013
- National AP Scholar August 2012
- REHSS High School Research Internship Acceptee (1/30 students nationwide) June 2012 - August 2012
- National Merit Commended Scholar December 2011