

Neeraj Varshney

Senior Applied Scientist, Amazon
PhD in Computer Science

nvarshney97@gmail.com
[Website](#) | [Google Scholar](#)
[LinkedIn](#)

RESEARCH SUMMARY

Senior Applied Scientist at Amazon working across the full foundation model development lifecycle — pretraining, mid-training, post-training, and reinforcement learning — with a focus on data-centric AI, reasoning, and Mixture-of-Experts architectures. Recipient of an organization-wide internal performance award for high-impact contributions to foundation-model training data quality and model performance.

My long-term research vision is to make foundation models learn more effectively from data, feedback, and interaction, rather than relying solely on increasing model size or compute. I pursue this through scalable training methodologies, teacher-guided RL, and taxonomy-driven data curation, validated across large-scale foundation models.

PhD in Computer Science from Arizona State University, advised by Prof. Chitta Baral. 15+ Publications at top-tier venues including ACL, EMNLP, NAACL, and EACL with 2,600+ citations on efficient, reliable, and trustworthy language models.

CORE EXPERTISE

Research Areas:

Pretraining, Mid-Training & Post-Training	Reinforcement Learning & Alignment
Data-Centric AI & Data Mixture Optimization	Synthetic Data Generation
Mixture-of-Experts (MoE)	Scaling Laws & Model Evaluation
LLM Reliability, Hallucination & Abstention	Agentic Systems & Tool Use

Technical Skills:

Languages:	Python, C++, SQL, Bash
ML / DL:	PyTorch, PyTorch Lightning, Hugging Face Transformers, TRL
Distributed Training:	Slime, Megatron-LM
Inference / Serving:	vLLM, SGLang
Data Processing:	PySpark, Hugging Face Datasets, Pandas
Experiment Tracking:	Weights & Biases, MLflow

EXPERIENCE

Amazon Science <i>Senior Applied Scientist</i>	2024 – Present <i>Palo Alto, California</i>
- Core contributor to three generations of large-scale Mixture-of-Experts foundation models trained on 40T+ tokens, ranging up to 800B+ total parameters. Contributions span pretraining data curation, multi-dimensional taxonomy filtering, evaluation, and reinforcement learning. - Owned the data curation and quality assessment stack end-to-end – quality filtering, deduplication, topic modeling, taxonomy development, and data-mixture optimization. This stack enabled models to outperform highest quality public datasets including Nemotron-CC, FineWeb and DCLM across numerous knowledge, reasoning, and coding benchmarks. - Designed and developed a multi-dimensional taxonomy curation framework spanning 14 orthogonal document-quality dimensions. Resulting filters recover high-value content from deprioritized web data tiers and surpass even the top-tier data on reasoning and coding benchmarks while extending the high-quality data supply by a full training stage without repetition. - Led improvements across code, math, and multilingual data tracks – including code-specific quality classifiers and synthetic data generation for reasoning. - Developing reinforcement learning approaches for improving reasoning and learnability of foundation models, leveraging privileged-information and teacher-guided training signals to generate solutions that are both correct and easier for smaller models to imitate. - Received organization-wide internal performance award for high-impact contributions to foundation-model training data quality and model performance.	

Tencent AI
NLP Research Intern

May 2023 – Aug 2023
Bellevue, Washington

- Built a hallucination mitigation framework for LLMs that identifies low-confidence generations, validates factual correctness using external evidence, and intervenes during decoding to mitigate them.
- Demonstrated effectiveness across long-form generation and question-answering tasks, improving factual reliability while preserving generation quality. This active intervention approach not only mitigates current hallucination but also prevents its propagation during autoregressive generation.

Amazon Science
Applied Scientist Intern, Alexa AI

May 2022 – Aug 2022
Seattle, Washington

- Developed efficient open-domain question answering systems combining parametric knowledge and retrieval-based reasoning.

Microsoft
Software Developer

July 2018 – Aug 2019
Bangalore, India

- Developed machine-learning-driven recommendation systems for Microsoft Teams to improve conversation discovery and user engagement.
- Built and deployed production services supporting large-scale enterprise users.

Samsung R&D Institute
Research Intern

Summer 2017
Bangalore, India

- Developed a 'context prediction' application leveraging event features such as app usage, location, and sensor data.

SELECTED PAPERS

Full publication list: [Google Scholar](#)

(2,600+ citations)

FOUNDATION MODELS & DATA-CENTRIC AI

1. Unlocking Latent Value: Taxonomy-Guided Recovery of High-Performing Data from Low-Tier Web Corpora Preprint

Neeraj Varshney, Sanket Lokegaonkar, Nasser Zalmout, Qingyu Yin, Priyanka Nigam, Bing Yin

- Proposed a taxonomy-driven framework for web data curation that models multiple semantic dimensions of document quality, including novel dimensions such as timeliness and cultural specificity.
- Demonstrated that taxonomy-filtered lower-tier web data can outperform conventional high-quality corpora on reasoning, coding, and knowledge benchmarks, highlighting the importance of data-centric approaches to foundation model training.

2. Expert Upcycling: Shifting the Compute-Efficient Frontier of Mixture-of-Experts Preprint

Chaitanya Dwivedi, Binxuan Huang, Himanshu Gupta, Pratik Jayarao, Neeraj Varshney, Bing Yin

- Introduced Expert Upcycling, a compute-efficient approach for expanding Mixture-of-Experts (MoE) models during continued pretraining by selectively duplicating and specializing high-utility experts.
- Showed that expert upcycling achieves comparable or superior model quality to training larger MoE models from scratch while significantly reducing training compute requirements.

REINFORCEMENT LEARNING & REASONING

3. Teacher-Guided Reinforcement Learning with Privileged Information In Progress, 2026

- Studying reinforcement learning approaches that incorporate privileged-information and teacher-guided signals to shape reasoning trajectories that are both correct and easier for smaller student models to imitate.
- Aimed at improving the learnability of foundation-model reasoning, bridging RL post-training and knowledge distillation to reduce the gap between teacher-only and student-deployable reasoning capability.

4. Stabilizing RL for Honesty Alignment in Language Models on Deductive Reasoning Preprint

Jiarui Liu, Kaustubh Dhole, Yingheng Wang, Haoyang Wen, Sarah Zhang, Haitao Mao, Gaotang Li, Neeraj Varshney, Jingguo Liu, Xiaoman Pan

- Studied faithful deductive reasoning in language models, focusing on scenarios where models should derive conclusions strictly from provided premises or abstain when evidence is insufficient.
- Introduced ACNCHOR, a reinforcement-learning-based training method that improves both reasoning accuracy and honest abstention behavior, resulting in more reliable and trustworthy reasoning systems.

LLM RELIABILITY & HALLUCINATION

5. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Preprint

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, Dong Yu

- Addressed the critical problem pertaining to hallucinations of LLMs and developed an approach that actively detects and mitigates hallucinations during the generation process.
- The method first identifies the candidates of potential hallucination leveraging the model’s logit output values, checks their correctness through a validation procedure, mitigates the detected hallucinations via prompting, and then continues generating the subsequent sentences.
- Showed the effectiveness of our approach in mitigating hallucinations of models such as GPT-3.5 and Vicuna in multiple tasks, such as article generation, multi-hop QA, and false premise QA.

6. Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA

ACL, 2023

Neeraj Varshney, Chitta Baral

- Introduced and mathematically formulated the “post-abstention” task: enabling models that abstain from low-confidence answers to re-attempt those instances, thereby increasing system coverage without significantly reducing accuracy.
- Developed Re-Examining the top-N Predictions (REToP) post-abstention recovery method, showing that even when the primary model abstains, the ground-truth answer often appears within its top-N logits, enabling reliable re-attempts using an auxiliary model and reducing system risk across IID and OOD QA benchmarks.

7. The Art of Defending: A Systematic Evaluation and Analysis of LLM Defense Strategies on Safety and Over-Defensiveness

ACL, 2024

Neeraj Varshney, Pavel Dolin, Agastya Seth, Chitta Baral

- Designed a large-scale evaluation framework encompassing safe and unsafe prompts to evaluate defense strategies across multiple state-of-the-art LLMs, addressing the research gap of disjoint evaluation suites.
- Demonstrated that existing LLM defense mechanisms suffer from over-defensiveness (i.e., rejecting safe inputs) and proposed new metrics and evaluation protocols to expose this issue.
- Quantified and compared the performance of diverse defense approaches on safety vs. over-defensiveness, yielding actionable findings and a publicly released dataset to enable future trustworthiness research.

8. Chaos with Keywords: Exposing Large Language Models Sycophancy to Misleading Keywords and Evaluating Defense Strategies

ACL, 2024

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, Chitta Baral

9. Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings

ACL, 2022

Neeraj Varshney, Swaroop Mishra, Chitta Baral

EFFICIENT INFERENCE & MODEL CASCADING

10. Investigating Acceleration of LLaMA Inference by Enabling Intermediate Layer Decoding via Instruction Tuning with ‘LITE’

NAACL, 2024

Neeraj Varshney, Agneet Chatterjee, Mihir Parmar, Chitta Baral

- Introduced the LITE framework, which applies intermediate-layer instruction-tuning in LLM so that earlier layers acquire generation capability — enabling token-level early exits without degrading output quality.
- Demonstrated dynamic, confidence-based early exiting at the token level, achieving 37.9% inference cost reduction for the 7B model and 46.4% for the 13B model, while maintaining comparable generation quality.
- Provided a deep analysis of layer-wise generation behavior and efficiency trade-offs, establishing a path toward scalable, cost-efficient LLM deployment in production settings.

11. Model Cascading: Towards Jointly Improving Efficiency and Accuracy of NLP Systems

EMNLP, 2022

Neeraj Varshney, Chitta Baral

- Introduced a model-cascading framework that dynamically routes inputs through a hierarchy of models of varying capacities, enabling easy instances to be handled by smaller, cheaper models and reserving large models for harder cases.
- Demonstrated that cascading can reduce computation cost by up to 88.9% (in a 3-model cascade setup) while simultaneously achieving up to 2.18% higher accuracy compared to using a single large model.
- Conducted comprehensive experiments across multiple NLP tasks and cascade depths (K-model settings), analyzing how introducing additional cascade stages further boosts efficiency, thereby paving the way for more deployable, cost-effective NLP systems.

EVALUATION & BENCHMARKING

12. ILDAE: Instance-Level Difficulty Analysis of Evaluation Data

ACL, 2022

Neeraj Varshney, Swaroop Mishra, Chitta Baral

- Devised a novel instance-level difficulty scoring framework across 23 diverse NLP datasets, enabling fine-grained assessment of how “hard” each test instance is for models
- Demonstrated five practical applications of our difficulty scores such as:
 - Conducting efficient-yet-accurate evaluations with fewer instances saving computational cost and time,
 - Improving the quality of existing evaluation datasets by repairing erroneous and trivial instances,
 - Indicating Out-of-Domain performance more reliably.

13. NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks

ACL, 2022

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, Ashwin Kalyan

- Designed and released NumGLUE, a multi-task benchmark of eight arithmetic reasoning tasks embedded in natural language, showing that even state-of-the-art LLMs underperform humans by 46%.
- Demonstrated that knowledge-retrieval multi-task learning method yields a gain of 3.4% per task compared to separate training—highlighting the value of cross-task knowledge sharing for numerical reasoning.
- This work serves as a foundational infrastructure for future research on “robust and general arithmetic reasoning” within language models, paving the way for more complex reasoning systems.

14. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks

EMNLP, 2022

Yizhong Wang, ..., Neeraj Varshney, ..., Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, Daniel Khashabi

- Introduced Super-Natural Instructions, a benchmark of 1,616 diverse NLP tasks and their expert-written instructions to evaluate instruction-based generalization.
- Developed and evaluated the “Tk-Instruct” model, demonstrating > 9% absolute performance gain over existing instruction-following models (e.g., InstructGPT) despite being an order-of-magnitude smaller
- Analysed scaling behaviour of instruction-tuned models — how number of seen tasks, examples per task, and model size affect cross-task generalization — providing practical roadmap for building more general-purpose NLP models.

15. LogicAttack: Adversarial Attacks for Evaluating Logical Consistency of Natural Language Inference

EMNLP, 2023

Mutsumi Nakamura, Santosh Mashetty, Mihir Parmar, Neeraj Varshney, Chitta Baral

16. “John is 50 years old, can his son be 65?” Evaluating NLP Models’ Understanding of Feasibility

EACL, 2023

Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Pal, S. Sawant, K. Scaria, S. Goyal, Chitta Baral

17. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models

ACL, 2024

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh, Arindam Mitra, Chitta Baral

18. Multi-LogiEval: Towards Evaluating Multi-Step Logical Reasoning Ability of Large Language Models

EMNLP, 2024

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, Chitta Baral

BOOKS

Advances in Multimodal Information Retrieval and Generation

Springer, 2024

Man Luo, Tejas Gokhale, Neeraj Varshney, Yezhou Yang, Chitta Baral

Synthesis Lectures on Computer Vision (SLCV)

- A comprehensive treatment of Transformer-based multimodal retrieval, generation, and retrieval-augmented generation across vision and language, covering foundational methods, architectures, and emerging research directions for students, researchers, and practitioners.

EDUCATION

Arizona State University

Tempe, AZ, USA

Ph.D. in Computer Science

2019 – 2024

- **Advisor:** Dr. Chitta Baral
- **CGPA:** 4.0/4.0
- **Outstanding CS PhD Graduating student Award** for the 2023-2024

BITS Pilani, Pilani Campus, India

Pilani, India

B.E (Hons) Computer Science

2014-2018

- **CGPA:** 9.11/10 (with Distinction)
- **Experience:** ‘Web Intelligence & Social Computing’ research lab under Prof. Poonam Goyal, CEERI research lab under Dr. J.L. Raheja.

HONORS AND AWARDS

Industry

- **Amazon Internal Performance Award** – for high-impact contributions to foundation-model training data quality and downstream model performance

Academic

- **Outstanding CS PhD Graduating student Award** for the 2023-2024 at Arizona State University
- **Outstanding Reviewer** for EACL'23 (Question Answering track).
- **Outstanding Research Award**, GPSA ASU, 2023
- SCAI **Doctoral Fellowship**, ASU, 2023.
- ASU **Jumpstart Research Grant**, 2023 and 2024.

SERVICE

Reviewing & Editorial

- **Area Chair**, ACL Rolling Reviews
- **Reviewer**: ACL, EMNLP, EACL (Outstanding Reviewer), Computational Linguistics Journal, COLM, AMLC, CVPR Workshop

Community & Outreach

- Author of 20+ ML/NLP articles on Medium with 100K+ cumulative views.
- Mentored several PhD interns and supported multiple co-authored publications.
- Math teacher for underprivileged students through the National Service Scheme (NSS), India.