

DEVANG KULSHRESHTHA

+1 929 394 0062 ◇ devangkulshreshtha12@gmail.com

Work Experience

Cruise

Senior Machine Learning Engineer

Sep 2026–Present

- Training perception models for improving self-driving cars using foundational models.

Amazon

Applied Scientist II, AWS AgentCore, Bedrock

Oct 2023–Sep 2026

- Lead the science direction for agent governance in AWS AgentCore, driving AgentCore Policy Engine.
- Mentor scientists and interns across the org and represent the team with external research partners.
- Set technical direction for AI Safety in Bedrock Guardrails, covering prompt-attack defense, harmful-content detection, and automated red-teaming; grew the business 800%+ in a year.
- Introduced self-jailbreaking, a red-teaming method in which the model guides its own compromise, now used to stress-test frontier models across the org.
- Published 5 papers at top NLP venues, filed 4 US patents, and mentored 3 interns to publication.

Applied Scientist I, AWS Bedrock, lex

Jul 2022–Sept 2023

- Served as science primary for ASR personalization and contextual biasing production work.
- Co-invented a retrieval-augmented inference architecture, enabling 10x entity catalogs at lower latency.

Applied Scientist Intern, Alexa

May 2021–Sept 2021

- Improved long-tail model accuracy by 12% via LLM-rescoring for European locales.

Machine Learning Engineer, Amazon Optimus

2018–2020

- Built production anomaly detection systems for large-scale invoice monitoring, reducing false positives by 30%.

Korbit.AI

2020–2022

Research Engineer

- Led research and deployment of question generation and adaptive feedback models powering a production intelligent tutoring system used at scale.
- Co-authored papers at EMNLP and IJCAI on domain adaptation and few-shot QA for education.

INRIA Research Labs, University of Rennes

Summer 2018

Research Intern

- Researched subgraph mining methods for interpreting deep network failures, improving diagnostic accuracy of incorrect-output detection.

Education

New York University, *PhD (part-time) in Machine Learning*

2025–Present

McGill University, *MSc in Computer Science, GPA: 4.0/4.0*

2020–2022

Indian Institute of Technology Varanasi, *B.Tech in Computer Science, GPA: 9.11/10*

2014–2018

Publications

1. Break Me If You Can: Self-Jailbreaking of Aligned LLMs via Lexical Insertion Prompting
D. Kulshreshtha, H. Su, H. Wang, C. Hegde. **COLM 2026** [\[Paper\]](#)
2. STAC: When Innocent Tools Form Dangerous Chains to Jailbreak LLM Agents
J.J. Li, J. He, D. Kulshreshtha, X. Xian, Y. Zhang, H. Su, S. Swamy, Y. Qi. **EMNLP REALM Workshop 2026** [\[Paper\]](#)
3. Understanding Uncooperative Behaviors in LLM based Multi-Agent Systems
D. Kulshreshtha, W. Du, R. Jain, S. Doss, H. Su, S. Swamy, Y. Qi. **EACL 2026** [\[Paper\]](#)

4. Relational Priors as Convergence Pressure in LLM-Based Multi-Agent Systems
M. Shen, C. Shang, S. Shahriar, [D. Kulshreshtha](#), Y. Zhang, S. Swamy, Y. Qi. **AACL 2026** [\[Paper\]](#)
5. Sequential Editing for Lifelong Training of Speech Recognition Models
[D. Kulshreshtha](#), S. Dingliwal, B. Houston, N. Pappas, S. Ronanki. **INTERSPEECH 2024** [\[Paper\]](#)
6. Retrieve and Copy: Scaling ASR Personalization to Large Catalogs
SM. Jayanthi, [D. Kulshreshtha](#), S. Dingliwal, S. Ronanki, S. Bodapati. **EMNLP 2023** [\[Paper\]](#)
7. Generalized Zero-Shot Audio-to-Intent Classification
V.R. Elluru, [D. Kulshreshtha](#), R. Paturi, S. Bodapati, S. Ronanki. **IEEE ASRU 2023** [\[Paper\]](#)
8. Multilingual contextual adapters for custom word recognition in low-resource languages
[D. Kulshreshtha](#), S. Dingliwal, B. Houston, S. Bodapati. **INTERSPEECH 2023** [\[Paper\]](#)
9. Improving Domain-Adaptive Generalization of CTC-Based ASR with Internal LM Estimation
N. Das, M. Sunkara, S. Bodapati, J. Cai, [D. Kulshreshtha](#), J. Farris, K. Kirchhoff. **ICASSP 2023** [\[Paper\]](#)
10. Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval
[D. Kulshreshtha](#), R. Belfer, I.V. Serban, S. Reddy. **EMNLP 2021** [\[Paper\]](#)
11. Few-shot question generation for personalized feedback in intelligent tutoring systems
[D. Kulshreshtha](#), M. Shayan, R. Belfer, S. Reddy, IV. Serban, E. Kochmar. **PAIS 2022** [\[Paper\]](#)
12. How Emotional Are You? Neural Architectures for Emotion Intensity Prediction in Microblogs
[D. Kulshreshtha](#), P. Goel, A.K. Singh. **COLING 2018** [\[Paper\]](#)
13. NLPRL-IITBHU at SemEval-2018 Task 3: Combining Linguistic Features and Emoji Pre-trained CNN for Irony Detection
H. Rangwani, [D. Kulshreshtha](#), A.K. Singh. **SemEval @ NAACL 2018** [\[Paper\]](#)
14. Prayas at EmoInt 2017: An Ensemble of Deep Neural Architectures for Emotion Intensity Prediction in Tweets
P. Goel, [D. Kulshreshtha](#), P. Jain, K.K. Shukla. **WASSA @ EMNLP 2017** [\[Paper\]](#)

US Patents

1. Removing bias from speech recognition models using internal language model estimates.
N. Das, M.N. Sunkara, S.B. Bodapati, J. Cai, [D. Kulshreshtha](#), J. Farris, N. Aldridge. [\[Patent\]](#)
2. Automated evaluation of evidence mapping models.
[D. Kulshreshtha](#), S. Dingliwal, S.B. Bodapati, K. Kirchhoff, S. Handa. [\[Patent\]](#)
3. Medical conversation summarization style intelligence.
A. Elangovan, [D. Kulshreshtha](#), K. Kirchhoff, L. Xu, S. Handa, S.B. Bodapati. [\[Patent\]](#)
4. Augmenting automated speech recognition neural networks with scalable vocabularies.
[D. Kulshreshtha](#), S. Dingliwal, S.B. Bodapati, V.R. Elluru, A. Mishra, K. Kirchhoff. [\[Patent\]](#)
5. Sequential Editing for Lifelong Training of Speech Recognition Models. *(Filed)*
[D. Kulshreshtha](#), N. Pappas, S. Dingliwal, B. Houston, S. Ronanki, V.R. Elluru.
6. Multilingual Contextual Adapters for ASR Personalization of low-resource languages. *(Filed)*
[D. Kulshreshtha](#), S. Dingliwal, B. Houston, S. Bodapati, S. Ronanki, J. Farris, V. Govindan, K. Kirchhoff.

Awards & Honors

- MITACS accelerate research scholarship at McGill University (65,000 CAD for 2 years).
- Kharusi International Science Fellowship (5,000 CAD).
- Shared task winner and invited for an oral talk at WASSA EMNLP 2017, Copenhagen, Denmark.