Gaurav Arora

Applied Scientist - 2 @ IML, Amazon

☎ 9888552871 | ✉ contactgauravforwork@gmail.com | www.arggaurav.com | goru001 | gaurav-arora-23593220

Education

Georgia Institute of Technology
Atlanta, Georgia

Masters in Computer Science with specialization in Machine Learning (OMSCS) *2020 - 2022*

- My GPA was 3.8/4. I pursued my Masters remotely during Covid in addition to full-time job at Haptik and then at Amazon.

Punjab Engineering College
Chandigarh, India

Bachelor of Technology (B.Tech.), Computer Science and Engineering *2014 - 2018*

- My CGPA was 7.77/10. I had the **best JEE-Mains Rank** in class.

Experience

Amazon
Bengaluru

Applied Scientist - 2 *Apr 2021 - Present*

- Helped in building and launching Conversational Shopping Assistant (CSA) for Amazon (India) in the pre and post LLM world.
- Built multiple **NLU components (Intent, NER, Query Reformulation, Policy Engine), Safeguard service, Next Question Suggestion service, Clarification and Product QnA tools** for the assistant, owning them from research to production. My work resulted in improvements across coverage and precision of bot response up to 10%.
- Worked on improving representation of code-mixed data in Transformer models, paper for which was accepted in **ACL 2023**.
- Worked on improving multilingual representations in LLMs, paper for which is under review at EMNLP 2024.
- Also worked on forecasting problems using Deep Learning and built the demand forecasting system for predicting smartphones demand for Amazon India, improving the model performance by 20%.
- Published **3 oral** (<10% acceptance rate) and **2 poster** (<30% acceptance rate) papers at Amazon's internal ML conference (AMLC).

Haptik (part of Jio Platforms Limited)
Mumbai

Machine Learning Scientist - 2 *July 2019 - Apr 2021*

- Built the **Intent Detection** System for Haptik's NLU Engine, owning it from research to production.
- The performance of new system was bench-marked against 10 data-sets, including 3 open source datasets and it was **25% more accurate** than their previous system, with latency under **150ms**.
- We also benchmarked the new system against popular NLU platforms like Dialogflow, RASA, LUIS in the paper **HINT3: Raising the bar for Intent Detection in the Wild** and Haptik's system achieved similar performance as others.

Goldman Sachs
Bengaluru

Software Development Engineer - 1 *June 2018 - July 2019*

- Worked on **Analytics** for Desktop Assistant to bring to the fore key metrics about Product Health and User Behavior.

Researchshala
Chandigarh

Co-Founder *Nov. 2016 - Mar. 2018*

- Built an online platform to connect professors to research-interns, helping them with their research projects.

Publications

Towards Abstractive Knowledge Representations in LLMs
Under review @ EMNLP 2024

Gaurav Arora, Shreya Jain, Vaibhav Saxena, Srujana Merugu

- We evaluate LLMs' ability to reason using two foundational relationships: **equivalence** and **inheritance**, by introducing novel tasks and evaluation benchmarks spanning six languages.
- Highlight systemic gaps in LLMs by showing that current SOTA LLMs often produce conflicting answers to the same questions across languages in 17.3-57.5% cases and also violate inheritance constraints in up to 37.2% cases.
- We propose novel **Compositional Representations** where tokens are represented as composition of equivalent tokens across languages, with resulting conflict reduction up to -4.7%.
- This paper was also accepted in Amazon's internal ML conference (AMLC 2024) as an oral paper (<10% acceptance rate).

Intent Detection in the Age of LLMs

Under review @ EMNLP 2024

Gaurav Arora, Shreya Jain, Srujana Merugu

(Industry Track)

- We adapt 7 SOTA LLMs using adaptive in-context learning and chain-of-thought prompting for intent detection, and compare their performance with contrastively fine-tuned sentence transformer (SetFit) models to highlight prediction quality and latency tradeoff.
- We propose a hybrid system using uncertainty based routing strategy to combine the two approaches, achieving the best of both worlds (i.e. within 2% of native LLM accuracy with 50% less latency).
- Introduce a two-step approach utilizing internal LLM representations, demonstrating empirical gains in OOS detection accuracy and F1-score by >5% for the Mistral-7B model.

MuST: A Multi Stage Targeting Framework for Promotional Campaigns

Accepted @ AMLC 2024

Apoorva Singh, Prince Jain, **Gaurav Arora**, Vivek Sembium

- In this work we propose a multi-stage framework for personalized targeting of campaigns in presence of marketing constraints. The proposed system achieved multi million dollar OPS impact over heuristic targeting.
- This paper was accepted in Amazon's internal ML conference (AMLC 2023) as a poster (<30% acceptance rate).

CoMix: Guide transformers to code-mix using POS structure and phonetics

Accepted @ ACL 2023 (Findings)

Gaurav Arora, Srujana Merugu, Vivek Sembium

- We propose CoMix, a pre-training approach to improve representation of code-mixed data in transformer models by incorporating phonetic signals, a modified attention mechanism, and weak supervision guided generation by parts-of-speech constraints.
- CoMix improves performance across four code-mixed tasks: machine translation, sequence classification, named entity recognition (NER), and abstractive summarization. It also achieves new SOTA performance for English-Hinglish translation and NER on LINCE Leaderboard and provides better generalization on out-of-domain translation.
- We also propose a new family of metrics based on phonetics and demonstrate that the phonetic variant of BLEU correlates better with human judgement than BLEU on code-mixed text.
- This paper was also accepted in Amazon's internal ML conference (AMLC 2022) as an oral paper (<10% acceptance rate).

Conversational Shopping Assistant: Bridging the gap between online and offline purchase experiences

Accepted @ AMLC 2023

Gaurav Arora, Pankaj Kumar, et al.

- This work describes how we built conversational shopping assistant for Amazon (India) which was launched in the IN marketplace for Laptop, Smartphone, and TV categories leading to multi-million dollar incremental OPS.
- In this paper we demonstrate how a conversational system can be bootstrapped successfully without the need for large volumes of chat data and evolved to utilize large language models (LLMs).
- This paper was accepted in Amazon's internal ML conference (AMLC 2023) as a poster (<30% acceptance rate).

DARE: Deep Affordability Recommendation Engine for ART events

Accepted @ AMLC 2022

Prince Jain, **Gaurav Arora**, Mohammed Abdulla, Vivek Sembium

- In this work we present a deep learning based solution which recommends the best affordability combination for each smartphone to minimize cost while achieving the ART (Amazon's sale events) event GMS goals.
- This paper was accepted in Amazon's internal ML conference (AMLC) in 2022 as an oral paper (<10% acceptance rate).

Inclusive Speech Detection using Pretrained Language Models

Accepted @ EACL 2021 LT-EDI workshop

Gaurav Arora*, Megha Sharma*

- In this paper we describe our system that ranked first in Hope Speech Detection (HSD) shared task and fourth in Offensive Language Identification (OLI) shared task, both in Tamil language.

Natural Language Toolkit for Indic Languages - iNLTK

Accepted @ EMNLP 2020 NLP-OSS workshop

Gaurav Arora

goru001/inltk

- iNLTK provides out-of-the-box support for Data Augmentation, Textual Similarity, Sentence Embeddings, Word Embeddings, Tokenization and Text Generation in 13 Indic Languages.
- iNLTK has 175,000+ downloads, 800+ stars on GitHub, and has been widely shared and appreciated on Twitter, LinkedIn, Reddit.
- By using pre-trained models and data augmentation from iNLTK, we achieve more than 95% of the previous best performance by using less than 10% of the training data on publicly available classification datasets.

HINT3: Raising the bar for Intent Detection in the Wild

Accepted @ EMNLP 2020 Insights workshop

GAURAV ARORA, CHIRAG JAIN, MANAS CHATURVEDI, KRUPAL MODI

hellohaptik/HINT3

- This paper introduces 3 new datasets created from real user queries received by live chatbots, evaluates popular NLU platforms and highlights critical gaps in language understanding.
- Paper introduces novel subset approach for evaluation and discusses trade-off between out-of-domain vs in-scope performance.

Pre-training ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection

Accepted @ FIRE 2020

GAURAV ARORA

goru001/nlp-for-manglish,nlp-for-tanglish

- This paper proposes pre-training ULMFiT on synthetically generated code-mixed data, generated by modelling code-mixed data generation as a Markov process.
- Model achieved 0.88 weighted F1-score for code-mixed Tamil-English language in Sub-task B and got 2nd rank on the leader-board.

Honors & Awards

Mar. 2021	**Indian Achievers Award 2020**, from Indian Achiever's Forum (IAF) in the Young Achievers Category for contribution in nation building through iNLTK	*India*
Mar. 2020	**Fast.ai International Fellow**, for contributions to Fast.ai forums	*Worldwide*
Mar. 2019	**Fast.ai International Fellow**, for contributions to Fast.ai forums	*Worldwide*
Dec.2018	**Top-17% , Human Protein Atlas Image Classification, Kaggle**, for developing Deep Learning model which classified mixed patterns of proteins in microscope images. The competition had 2172 teams, but I participated individually and hence had 100% contribution in the 366th placed solution	*Worldwide*
Oct.2017	**1st Prize,IEEE-Hackathon**, for developing chat-bot to help people with emotional decisions in life	*PEC*
Feb.2016	**Top-100**, among 500,000 students in IT-Olympiad,2016.	*India*
Oct.2016	**2nd-Prize, IEEE-Hackathon**, for developing an Augmented reality application to help teachers	*PEC*
Mar.2016	**All India Rank-6**, in IEEE Programming League, among over 1200 undergraduate students.	*India*
Mar.2016	**2nd Rank**, CodeWars, a competitive-programming event hosted by IEEE,PEC on CodeChef	*PEC*
Nov.2016	**Research Scholarship**, for Personal Emotional Doctor - Bot	*PEC*
2014	**All India Rank-885**, JEE-Mains, among 1.4 million candidates	*India*
2014	**1st Rank-Opener,PEC**, for **best** JEE-Mains rank among 600 students of the session 2014-2018.	*PEC*
2014	**1 Lakh Scholarship**, from CBSE for 96.4% marks in 12th Boards and 10 CGPA in 10th.	*India*
2014	**Letter of Appreciation**, from HRD Ministry,Govt. of India for 96.4% in CBSE-12th exams	*India*
2011	**Catch Them Young**, Was among the **top-40** students selected from tricity by INFOSYS for 2-week Programming-Basics training on their campus	*INFOSYS*

Skills & Courses

Mathematics	Discrete Structures for Computer Science, Bayesian Statistics, Vector Calculus, Operations Research, Mathematics for Machine Learning: Linear Algebra and Multivariate Calculus
Computer Science	Data Structures and Algorithms, Intro to Graduate Algorithms, Computer Architecture and Organization, OOP, Microprocessor, DBMS, Operating Systems, Computer Networks, Theory of Computation, Artificial Intelligence, Knowledge Based AI, Machine Learning, Deep Learning, Reinforcement Learning, Computer Vision, Big Data for Healthcare, Machine Learning for Trading, Computer Graphics, Mobile Computing
Programming & Web	Python, C, C++, Javascript, TypeScript, EcmaScript6, AngularJS, ReactJS, Angular4, Webpack
Machine Learning tools	Pytorch, Pandas, Numpy, ScikitLearn, SciPy, Fastai, Transformers library

Extracurricular Activity

MOTIVATIONAL-SPEAKER *2014 - PRESENT*

- Given more than 10 motivational lectures at schools, college etc to an audience of about 100, talking about *student-life challenges* and *success-hacks*.