Harsh **Kohli**

Graduate Research Assistant

Columbus, OH, USA

+1-614-554-9384 | harsh14791@gmail.com | HarshKohli | harsh-s-kohli | Harsh Kohli

Education

The Ohio State University — *Columbus, USA*

PhD Computer Science & Engineering — *Aug. 2022 - Present*

- Language Models & Natural Language Processing

Georgia Institute of Technology — *Atlanta, USA*

M.S Computer Science — *Jan. 2018 - May 2020*

- Specialization in Machine Learning

Birla Institute of Technology & Science, Pilani - K.K Birla Goa Campus — *Goa, India*

B.E (Hons.) Computer Science — *Aug. 2010 - May 2014*

Work Experience

Meta — *Menlo Park (CA), USA*

Research Scientist Intern — *Jun 2025 - Aug 2025*

- Worked with the GenAI Monetization team to benchmark and optimize LLM agents for business messaging
- Analyzed the impact of a thinking budget on different reasoning models for tasks of varying complexity
- Implemented a token-budget aware reasoning strategy to adaptively adjust the number of reasoning tokens based on input prompts

Amazon — *Cambridge (MA), USA*

Applied Scientist Intern — *May - Aug 2024*

- Part of the Amazon AGI LLM Foundations team working on Supervised Finetuning of Large Language Models
- Developed EWoRA - a method to improve low-rank adaptation of language models for heterogenous finetuning data using exert weighting. The work is accepted at IJCNLP-AACL 2026 - Findings Track

Amazon — *Cambridge (MA), USA*

Applied Scientist Intern — *May - Aug 2023*

- Worked with the Amazon Alexa Entity Resolution team.
- Developed a method for incorporating additional word and corpus-level features to improve state-of-the-art performance on Bilingual Lexicon Induction. The work is accepted at NAACL 2024 - Findings Track

Compass — *Bangalore, India*

Senior Machine Learning Scientist 2 — *Mar. 2021 - Apr. 2022*

- Implemented a low latency partial-query auto-suggest system incorporating both textual and non-textual features for real-estate search
- Worked on creating a unified representation using the multi-modal features on a real-estate listing page
- Developed a task-oriented chatbot to answer consumer queries in the absence of an agent - trained the dialogue state tracker to recognize the important entities and detect user intent as well as imbued the bot with a Reading Comprehension model to answer questions from text

Salesken — *Bangalore, India*

Technical Architect - Data Science — *Jan. 2020 - Mar. 2021*

- Involved in the company's ideation, planning and strategy to leverage NLP techniques to improve sales conversations
- Led and mentored a team of Data Scientists to build several of the products key components including models for semantic matching and inference, emotion detection, dialogue state tracking, and fast semantic vector search
- Optimized the various Deep Learning models for latency and throughput, and set up the pipeline to drive them to production - Docker containerization, half-precision GPU inference, and auto-scaling using Kubernetes clusters

Microsoft — *Hyderabad, India*

Data Scientist 2 — *Feb. 2018 - Jan. 2020*

- Developed an architecture for Mixed Objective, Block level optimization for web search recommendation in Bing using a Pointing Decoder model and reinforcement learning objectives - the research was accepted at SIGIR 2020
- Worked on Bing's domain-specific passage ranking and Machine Comprehension abilities. As part of the work, we published a novel lightweight Attention-LSTM model for online ranking through a paper at CIKM 2018
- Session-context aware and Multi-Task DL systems for follow up query suggestion in Bing's Related Questions feature
- Entity disambiguation using Bi-Encoders in Web Queries using various negative-sampling strategies and optimization techniques

IPsoft *Bangalore, India*

R&D Engineer *Jun. 2014 - Feb. 2018*

- Worked on several modules including Machine Comprehension, Dialogue Management, and Translation, for Ipsoft's virtual assistant 'Amelia'
- Applied traditional logic and rule-based approaches to solve for these problems using taxonomies, ontologies and various other linguistic resources such as WordNet, FrameNet, PropBank etc.
- Led a team working on improving state of the art approaches on tasks like Question Answering, Natural Language Inference etc. using Attention-Matching sequence models under the guidance of Prof. Manning

Select Research Publications

Loop, Think, & Generalize: Implicit Reasoning in Recurrent-Depth Transformers
Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, Yuekun Yao
Preprint. Under review.
doi:10.48550/...822

EWoRA: Expert Weighted Low-Rank Adaptation for Heterogeneous Data
Harsh Kohli, Helian Feng, Lenon Minorics, Bhoomit Vasani, Xin He, Ali Kebarighotbi
Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL), 2025
doi:10.18653/...108

GroundCocoa: A Benchmark for Evaluating Compositional & Conditional Reasoning in Language Models
Harsh Kohli, Sachin Kumar, Huan Sun
Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), 2025
doi:10.18653/...420

How Lexical is Bilingual Lexicon Induction?
Harsh Kohli, Helian Feng, Nicholas Dronen, Calvin McCarter, Sina Moeini, Ali Kebarighotbi
Findings of the Association for Computational Linguistics: NAACL 2024, 2024
doi:10.18653/...273

Training Bi-Encoders for Word Sense Disambiguation
Harsh Kohli
Document Analysis and Recognition – ICDAR 2021, Springer International Publishing, 2021
doi:10.1007/..._53

Transfer Learning and Augmentation for Word Sense Disambiguation
Harsh Kohli
Advances in Information Retrieval - ECIR 2021, Springer International Publishing, 2021
doi:10.1007/..._29

Training Mixed-Objective Pointing Decoders for Block-Level Optimization in Search Recommendation
Harsh Kohli
Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020
doi:10.1145/...236

AQuPR: Attention Based Query Passage Retrieval
Parth Pathak, Mithun Das Gupta, Niranjan Nayak, Harsh Kohli
Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM), 2018
doi:10.1145/...323

Document Categorization using Semantic Relatedness & Anaphora Resolution: A Discussion
Kaustubh Dhole, Harsh Kohli
2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2015
doi:10.1109/...279

Optimal Route Searching in Networks with Dynamic Weights Using Flow Algorithms
Sunit Singh, Ram Prasad Joshi, Harsh Kohli
2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015
doi:10.1109/...37

For a complete list of publications, visit **Google Scholar**

Select Academic Projects

LLM Digital Twins

- Collaborating with Prof. Ryan Kennedy to develop a digital twin for political survey participants and model their responses.
- Designing prompting and fine-tuning methods to generate accurate survey responses based on respondent personas.

Jack Watson Research Squad for Plagiarism Detection

- Working under Prof. Thad Starner (GeorgiaTech) towards developing a Chatbot to catch plagiarism on homework-for-hire websites.
- Developed modules for plagiarism intent detection, homework similarity-matching algorithms, general conversation through Seq2Seq models, text auto-correct, and keyword extraction. The project was accepted as a Work-in-Progress paper at the Learning@Scale conference 2019.

Zero-Shot Retrieval on Tabular Data

- Proposed a method to encode a natural-language question and table information using features such as column-names and row content.
- Trained an LSTM-based deep learning model for Information Retrieval using this method over tabular data.