Harshit Joshi

josharshit@gmail.com
harshitj@stanford.edu
https://github.com/duskybomb
https://www.linkedin.com/in/harshitjos
https://scholar.google.com/citations?user=NFZwEmUAAAAJ

Education

- **Stanford University** *2023 - Present*
  Doctor of Philosophy (Ph.D) in Computer Science GPA: 4.1
- **Cluster Innovation Center, University of Delhi** *2017 - 2021*
  Bachelor of Technology (B. Tech) in Information Technology and Mathematical Innovations Percentage: 87.78%

Research Interests

Large Language Models, Reasoning with LLMs, Systems for Conversational Agents, ML for Code

Research Experience

- **Microsoft Research** *Nov 2021 - July 2023*
  Research Fellow (Predoctoral) with the PROSE group Bengaluru, India
  - Designed and trained a small language model for spreadsheets (60M) that outperforms much larger LLMs (175B) in formula repair and formula autocompletion. This work was covered by media outlets. (AAAI 24)
  - Created multilingual repair framework, RING, that leverages LLMs and compiler diagnostics. (AAAI 23)
  - Built neurosymbolic program repair framework for Excel and PowerApps. (In private preview) (OOPSLA 22).
- **Defence Research and Development Organisation, Govt. of India** *June 2019 - Oct 2019*
  Research Intern New Delhi, India
  - Worked with CityScape Dataset for Image Segmentation through the PyTorch implementation of DeepLabV3+.
  - Fine-tuned the Image Segmentation model for cognitive navigation and mapper.

Papers (*denotes equal contribution)

- **H. Joshi**, A. Ebenezer, J. Cambronero, S. Gulwani, A Kanade, V. Le, I. Radicek and G. Verbruggen. *FLAME: A small language model for spreadsheet formulas.* To be presented at Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38., 2024. **Oral Presentation (top ∼ 2%)**
- **H. Joshi**, J. Cambronero, S. Gulwani, V. Le, I. Radicek and G. Verbruggen. *Repair Is Nearly Generation: Multilingual Program Repair with LLMs.* Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 4. 2023
- R. Bhavishi*, **H. Joshi***, J. Cambronero, A. Fariha, S. Gulwani, V. Le, I. Radicek and Ashish Tiwari. *Neurosymbolic Repair for Low-Code Formula Languages.* In Proceedings of the ACM on Programming Languages (OOPSLA) 2022.
- R. Sawhney*, **H. Joshi***, A. Nobles*, and R. R. Shah. *Towards Emotion-and Time-Aware Classification of Tweets to Assist Human Moderation for Suicide Prevention.* In International AAAI Conference on Web and Social Media 2021.
- R. Sawhney*, **H. Joshi***, R. R. Shah, and L.Flek. *Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning.* In North American Chapter of the Association for Computational Linguistics 2021.
- R. Sawhney*, **H. Joshi***, L.Flek, and R. R. Shah. *Phase: Learning Emotional Phase-Aware Representations for Suicide Ideation Detection on Social Media.* In European Chapter of the Association for Computational Linguistics 2021.
- R. Sawhney, **H. Joshi**, S. Gandhi, D. Jin, and R. R. Shah. *Robust Suicide Risk Assessment on Social Media via Deep Adversarial Learning.* In Journal of the American Medical Informatics Association 2021.
- R. Sawhney,**H. Joshi**, S. Gandhi , and R. R. Shah. *Towards Ordinal Suicide Ideation Detection on Social Media.* In ACM International Conference on Web Search and Data Mining 2021.
- R. Sawhney,**H. Joshi**, S. Gandhi, and R. R. Shah. *A Time-aware Transformer based Model for Suicide Ideation Detection on Social Media.* In Conference on Empirical Methods in Natural Language Processing 2020.

Professional Experience

- **Arkifi.ai** *July 2023 - September 2023*
  AI/ML Research Consultant California, USA
  - Built the pipeline for table information extraction in spreadsheets increasing baseline performance by 30%
- **Supedio GmbH** *Jan 2021 - Nov 2021*
  Research Software Engineer Dresden, Germany
  - Developed the financial data extraction pipeline from digital documents, invoices and purchase orders.
  - Built automation tools for the sales team, reducing person-hours employed per day by 400%.
  - Built algorithmic pipelines for address deduplication, finding 18% duplicates in client's "gold" database.
- **Cronycle Ltd.** *Jan 2019 - July 2019*
  Software Engineering Intern New Delhi, India
  - Migrated batch jobs for retrieving RSS to real-time using Kafka and ElasticSearch, reducing latency by 5 min.
  - Increased RSS collection dump by 10% by identifying new data sources and processing them to MongoDB.

Technical Skills

Languages: Python, C++, Java, Javascript, C#, Go
Software and Tools: Pytorch, Langchain, MongoDB, Git, Elastic Search, PostgreSQL, Flask, Airflow, Kafka

Positions of Responsibility

- **Student Coordinator,** Delhi University Innovation Council *Oct 2018 - Sept 2019*
- **Lead Organizer,** Convoke 3.0: Technical Fest at University of Delhi *Aug 2019 - Oct 2019*
- **Head of External Affairs,** #Include: Computer Society CIC, DU *Aug 2018 - Aug 2019*

Achievements

- Selected for Oral Presentation at AAAI 2024 - FLAME: A small language model for spreadsheet formulas.
- Oral Presentation at WSDM (Virtual) 2021 - Towards Ordinal Suicide Ideation Detection on Social Media.
- Oral Presentation at EMNLP (Virtual) 2020 - A Time-aware Transformer based Model for Suicide Ideation Detection on Social Media.
- Received Honorable Mention at COMAP 2020 (Highest ranked Indian team).
- Summer Fellowship 2019 for Mathematical Finance Scholar - IAS, INSA, NAS, CMI
- Selected for Google Summer of Code 2018, Invoice2data library for extracting financial data
- ACM ICPC Regionals 2018 - Honourable Mention

Talks

- **OOPSLA** Virtual, 2022
- **NAACL** Virtual, 2021
- **EACL** Virtual, 2021
- **EMNLP** Virtual, 2020
- **PyData Delhi** 2019
- **ML Research, University of Delhi** 2018-2021

Services

- **Reviewing** AAAI, CA2MH @ ICML
- **Volunteer** EMNLP, EACL, ICWSM, AAAI