SHASHWAT SINGH

• shashwat.s@research.iiit.ac.in • Linkedin: Shashwat • Github: shashwat1002 • Scholar: Shashwat

EDUCATION

B. Tech. Computer Science with M. S. by Research in Computational Linguistics Duration: 2020-2025
International Institute of Information Technology Hyderabad 8.72/10.0 CGPA
Societies: Open Source Developers' Group (Lead), Debate Society (Lead)
Relevant coursework: Topics in Deep Learning, Advanced NLP, Maths of Generative Models, Information-Theory
Awards:

SELECTED EXPERIENCE

International Institute of Information Technology, Hyderabad: *Student Researcher* Feb 2024 - Present
Studying compositionality in CLIP with Dr. Makarand Tapaswi

- Developing a methodology to align pure text and pure image embeddings.
- Probing studies for compositionality in joint vision-language models - specifically CLIP.

Trexquant Business Consulting LLP, Gurgaon: *Quantitative Research intern* May 2024 - July 2024
Machine Learning models for medium frequency signals

- Designed and built a RAG based code-generation model.
- Engineered features for a gradient-boost model to make earning's predictions.

Indian Institute of Science, Bangalore: *Research Intern* May 2023 - Nov 2023
Controlled Generation in Large Language Models with Dr. Danish Pruthi and Dr. Ponnurangam Kumaraguru

- Designed and implemented experiments to study effects of model editing of Language Models using Knowledge Graphs.

International Institute of Information Technology, Hyderabad: *Student Researcher* May 2022 - **Present**
Investigating Vision Language models and negation in Large Language Models with Dr. Ponnurangam Kumaraguru

- Investigating Alignment of text-in-image capabilities in CLIP shared encoder.
- Worked on a methodology for interventions on pre-trained generative Language Models to elicit desirable properties in their generation. Application in Detoxification of pre-trained LMs.
- Showed using probing studies that negation information while encoded is not used to make factual evaluations.

International Institute of Information Technology, Hyderabad: *Student Researcher* Jan 2023 - May 2023
Investigating training dynamics using an information theoretic lens with Dr. Manish Shrivastava

- Designed toy tasks and ran training experiments so as to formulate information propagation in a model undergoing phases of training.

Google Summer of Code, Remote: *Open Source* Jun 2022 - Nov 2022
Project under ccextractor: Porting an OCR module

- Ported a C module for OCR on video files to Rust. Patched breaking changes for FFMPEG-5 compatibility.

PUBLICATIONS AND TALKS

- **Singh, S.***, Ravfogel, S.*, Herzig, J., Aharoni, R., Cotterell, R., Kumaraguru, P. (2024). **Representation Surgery: Theory and Practice of Affine Steering** [Poster] *International Conference on Machine Learning 2024 (**ICML**)*
- **Singh, S.***, Goel, S.*, Vaduguru, S., & Kumaraguru, P. (2023). **Probing Negation in Language Models** [Poster]. Appeared in *ACL 2023, 8th Workshop on Representation for Learning (**REPL4NLP**)*

AWARDS AND HONOURS

- Dean's Research Award
- Centre for AI Safety student research stipend program
- Dean's Merit List for three semesters (top 20%)

TEACHING EXPERIENCE

International Institute of Information Technology, Hyderabad: *Teaching Assistant* January 2024 - May 2024
Responsible and Safe AI Systems

- Designed a part of the course curriculum.
- Conducted tutorials on foundations of interpretability, and ML architectures.

Indian Institute of Science, Bangalore: *Teaching Assistant* Nov 2023
AI-mpower Programme

- Designed and conducted tutorials and lab content covering basics of ML, Deep Learning and NLP.
- Program for professors and students from universities across India to learn about Machine Learinng and AI.

International Institute of Information Technology, Hyderabad: *Teaching Assistant* January 2023 - May 2023
NLP,

- Conducted tutorials on Machine Learning in the context of NLP.

International Institute of Information Technology, Hyderabad: *Teaching Assistant* August 2023 - November 2023
Discrete Structures

- Conducted tutorials on the basics of theorem proving.

OPEN SOURCE

- Kiwix: merged changes that prevent redirect loops in their web archive format.
- SageMath: added a few doctests in their Linear Algebra module.

TECHNICAL SKILLS

Languages: Python, C, C++, Rust, SQL, SPARQL **Frameworks:** PyTorch, NLTK, Spacy, Django