

# Shashwat Singh

PHD STUDENT

NYU Center for Data Science

✉ ss20428@nyu.edu | 🏷 shashwat1002.github.io | 📡 shashwat1002

## Education

---

### NYU Centre for Data Science

PHD

New York

2025 - 2030 (expected)

- Advised by Dr. He He and Dr. Tal Linzen

### IIIT Hyderabad

#### B. TECH IN COMPUTER SCIENCE + M. S. BY RESEARCH IN COMPUTATIONAL LINGUISTICS

Hyderabad

2020 - 2025

- Advised by Dr. Ponnurangam Kumaraguru in the integrated bachelor's and masters by research program
- CGPA: 8.67 / 10.0

## Research Experience

---

### Emergence of Text Semantics in CLIP Image Encoders

IIIT HYDERABAD

Feb 2024 - Oct 2024

- Project in collaboration with Sreeram Vennam, Anirudh Govil, under the supervision of **Dr. Ponnurangam Kumaraguru**
- Found through controlled experiments that the Image Encoder in CLIP has some notion of textual semantics.

### Representation Surgery: Theory and Practice of Affine Steering

GOOGLE RESEARCH, ETH ZURICH, IIIT HYDERABAD

Dec 2023 - May 2024

- Project in collaboration with **Dr. Shauli Ravfogel**, **Dr. Ryan Cotterell**, and **Dr. Ponnurangam Kumaraguru**
- Formalized steering functions on pre-trained language modeling representations.
- Provided and solved optimization objectives for generating counterfactuals. Demonstrated results on controlled generation and debiasing.

### Investigating Text to Image Mapping in CLIP and Diffusion Models

IIIT HYDERABAD

Jan 2024 - Present

- Project in collaboration **Dr. Makarand Tapaswi**
- Investigating VAE latent space for discriminative properties.
- Developing a methodology to align pure text and pure image embedding.
- Probing studies for compositionality in joint vision-language models - specifically CLIP.

### Probing Negation in Language Models

LTI - CARNEGIE MELLON UNIVERSITY, IIIT HYDERABAD

Aug 2022 - May 2023

- Project in collaboration with **Shashwat Goel**, **Saujas Vaduguru** and advised by **Dr. Ponnurangam Kumaraguru**
- Showed using probing studies that negation information while encoded is not used to make factual evaluations.

## Publications

---

\* refers to equal contribution

**Shashwat Singh\***, Shauli Ravfogel\*, Jonathan Herzog, Roee Aharoni, Ryan Cotterell, Ponnurangam Kumaraguru. **Representation Surgery: Theory and Practice of Affine Steering** [Poster] *International Conference on Machine Learning 2024 (ICML)*

Sreeram Vennam\*, **Shashwat Singh\***, Anirudh Govil, Ponnurangam Kumaraguru **Emergence of text semantics in CLIP image encoders** [to appear as a poster] *NeurIPS 2024, 2nd workshop on Unifying Representations in Neural Models (UniReps, NeurIPS)*

**Shashwat Singh\***, Shashwat Goel\*, Saujas Vaduguru, Ponnurangam Kumaraguru. **Probing Negation in Language Models** [Poster]. Appeared in *ACL 2023, 8th Workshop on Representation for Learning (REPL4NLP, ACL)*

## Professional Experience

---

### Trexquant Business Consulting LLP

#### QUANTITATIVE RESEARCH INTERN

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

Gurgaon, India

May 2024 - July 2024

### Indian Institute of Science (IISc)

#### VISITING RESEARCH INTERN

- Hosted by **Dr. Danish Pruthi**
- Designed and implemented experiments to study effects of model editing of Language Models using Knowledge Graphs.
- investigated concept-level controlled generation

Bangalore, India

May 2023 - Nov 2023

### Google Summer of Code

#### OPEN SOURCE DEVELOPER

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

Remote

Jun 2022 - Nov 2022

## Teaching Experience and Talks

---

2024 **Responsible AI Systems**, NPTEL, Guest Lecturer

Remote

2024 **Responsible and Safe AI Systems**, IIIT Hyderabad, Teaching Assistant

Hyderabad

2023 **AI empower program**, Indian Institute of Science, Teaching Assistant

Bangalore

2023 **Introduction to NLP**, IIIT Hyderabad, Teaching Assistant

Hyderabad

2022 **Discrete Structures**, IIIT Hyderabad, Teaching Assistant

Hyderabad

## Awards, Fellowships, & Grants

---

2024 **Dean's Research Award**, IIIT Hyderabad

2022 **AI Safety Student Research Stipend**, Centre for AI Safety

2020, 2021 **Dean's Merit List (top 20%)**, IIIT Hyderabad

## Open Source

---

**Kiwix**, merged changes that prevent redirect loops in their web archive format.

**SageMath**, added a few doctests in their Linear Algebra module.

## Service

---

2024 **NeurIPS 2024, 2nd workshop on Unifying Representations in Neural Models**, Reviewer

2023 **ACL 2023, 8th Workshop on Representation for Learning**, Reviewer

## University positions

---

2022 - 2023 **Open Source Developer's Group @ IIIT Hyderabad**, Lead

2022 - 2023 **Debate Society @ IIIT Hyderabad**, Lead

2022 - 2023 **Theory Group @ IIIT Hyderabad**, Moderator

## Skills

---

Python, PyTorch, Huggingface, Django, C, C++, Rust, SQL, SPARQL