

Shashwat Singh

PHD STUDENT

NYU Center for Data Science

✉ ss20428@nyu.edu | 🏠 shashwat1002.github.io | 📷 shashwat1002

Education

NYU Centre for Data Science

PHD

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

IIIT Hyderabad

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

- Advised by Dr. Ponnurangam Kumaraguru in the integrated bachelor's and masters by research program
- Led University Societies: Opensource Developers' Group, Theory Group
- CGPA: 8.67 / 10.0

New York

August 2025 - August 2030

(expected)

Hyderabad

2020 - 2025

Publications

* refers to equal contribution

Shashwat Singh, Tal Linzen, Shauli Ravfogel. **Can LLMs Introspect? A Reality Check.** [Accepted in] *Conference on Language Modeling 2026 (COLM)*

Shashwat Singh*, Shauli Ravfogel*, Jonathan Herzig, Roei 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** [Poster, Accepted in] *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)*

Selected Research Projects

Investigating claims of introspection in Language Models

NYU CENTRE FOR DATA SCIENCE

Dec 2025 - July 2026

- Project in collaboration with Tal Linzen and Shauli Ravfogel
- We clearly operationalize necessary conditions to claim introspection in Language Models and show that some popular works don't meet these conditions. Accepted in COLM 2026.

Does Vision Language Alignment help the language decoder?

NYU CENTRE FOR DATA SCIENCE

Sept 2025 - Present

- Project in collaboration with He He, and Tal Linzen
- We investigate whether popular paradigms in vision-language alignment improve the language decoder. Developing an alignment paradigm that utilizes vision data to make better language models.

Emergence of Text Semantics in CLIP Image Encoders

IIIT HYDERABAD

Feb 2024 - Oct 2024

- Project in collaboration with Sreeram Vennam, Anirudh Govil, **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 - September 2024

- Project in collaboration **Dr. Makarand Tapaswi**
- Investigated VAE latent space for discriminative properties.
- Probing studies for compositionality in joint vision-language models - specifically CLIP. Found that VAEs in Stable Diffusion learn to be almost linear.

Professional Experience and Contract Work

UK AI Security Institute

Remote

CONTRACTUAL

Jan 2025

- Implemented an automated pipeline that operates on SAE features to detect collusion between agents. Implements Lermen & Dziemian, 2024.

Trexquant Business Consulting LLP

Gurgaon, India

QUANTITATIVE RESEARCH INTERN

May 2024 - July 2024

- 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 (IISc)

Bangalore, India

VISITING RESEARCH INTERN

May 2023 - Nov 2023

- 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

Google Summer of Code

Remote

OPEN SOURCE DEVELOPER

Jun 2022 - Nov 2022

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

Teaching Experience and Talks

2026	Large Language Models: Evaluation and Applications , NYU CDS, Section Leader	New York
2024	Responsible AI Systems , NPTEL, Guest Lecturer	Remote
2023	AI empower program , Indian Institute of Science, Teaching Assistant	Bangalore
2022-2025	Introduction to NLP, Responsible and Safe AI Systems, 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

Skills

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