Vedant Palit

vedantpalit@kgpian.iitkgp.ac.in | vedantpalit.github.io | github.com/vedantpalit | google scholar

Education

Indian Institute of Technology(IIT) Kharagpur, India 2021 – 2026
Btech in Industrial & Systems Engineering and Mtech in Financial Engineering

Birla High School 2008 – 2021
Subjects: *English, Physics, Chemistry, Mathematics, Computer Science*

Publications

Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP [**Link**]
Published at the ICCV 2023 Workshop on Closing The Loop Between Vision and Language - ***First Author***

Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust [**Link**]
Published and Presented at the IEEE Conference on Artificial Intelligence 2023

WellDunn: On the Robustness and Explainability of LMs and LLMs [**Link**]
Accepted at the ACL ARR 2024 and Under Review at EMNLP 2024

What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline [**Link**]
Under Review at EMNLP 2024

Research Experience

Mechanistic Interpretability of Vision-Language Models Providence, RI (Remote)
As a part of the Eickhoff AI Lab, Brown University *Feb 2024 – June 2024*

- Developed and created an extensive pipeline for an in-depth mechanistic intepretability study of the BLIP VL model through path patching and knockouts.
- Introduced the novel semantic minimal pair and symmetric text replacement corruption scheme demonstrating more reliable results from causal mediation analysis over the pre-existing gaussian noise corruption scheme.

QA-based Chunk Formatting for Retrieval Improvement Columbia, SC (Remote)
In Collaboration with Kaushik Roy Phd, University of South Carolina *March 2024 – June 2024*

- Implemented and benchmarked Vanilla and Sentence-Window RAG on the MultiHopRAG dataset, using multiple open-source models such as Llama2, OrcaMini-3B and evaluation metrics such as BLUE, ROUGE-L and NUBIA.
- Devised a novel paradigm of generating independent and closed context question-answer pairs to improve retrieval capability of vanilla RAG.

Causal Intervention on the BLIP Architecture Pittsburgh, PA (Remote)
In Collaboration with Rohan Pandey, Carnegie Mellon University *April 2023 – Sep 2023*

- Created a pipeline adapting causal mediation analysis to interpret blackbox architectures of VL transformers.
- Implemented the method on the BLIP transformer and used the COCO-QA dataset to study the effect of various layers on the final outputs.

Wellness Dimensions Benchmark for Explainability of LMs Baltimore, MD (Remote)
Under the guidance of Prof Manas Gaur, University of Maryland, Baltimore *Nov 2022 – Dec 2023*

- Trained various general and domain-specific models for suicide risk assessment, using the gamblers and cross entropy loss functions on annotated datasets containing social media posts classified into 6 different wellness dimensions.
- Utilised singular value decomposition to analyse the impact of the loss function on the attention scores of the models.

Knowledge Graph Guided Semantic Evaluation of LMs For User Trust Columbia, SC (Remote)
In Collaboration with Kaushik Roy Phd, University of South Carolina *Feb 2023 – March 2023*

- Developed a novel evaluation method to measure error in reconstruction of masked knowledge graph structures from outputs by LLMs.
- Analysed and benchmarked the performance of GPT-3.5, GPT-J and GPT-NeoX in reconstructing KG paths using the evaluation metric.

Relevant Coursework

- **University**: Regression and Time Series Models(MA60280), Machine Learning(AI41002) Transform Calculus(MA20202), Operations Research-I(IM21201), Programming and Data Structures(CS10003), Linear Algebra-Numerical and Complex Analysis(MA11004)
- **MOOCs**: Natural Language Processing with Deep Learning(CS224N), Introduction to Algorithms(MIT 6.006), Neural Networks and Deep Learning, Machine Learning

Technical Skills

Programming Languages: C/C++, Python, MATLAB
ML-DL: TensorFlow, Pytorch, Torchvision, Sklearn, Caffe
CV-NLP: Transformers, OpenCV, PIL, Llama-Index
Miscellaneous: Mysql, LaTeX, HTML, Markdown, Git

Awards and Achievements

JEE Advanced: Placed in the top 0.5% nationally among candidates appearing in JEE Advanced, 2021.
JEE Mains: Placed in the top 0.8% nationally among candidates appearing in JEE MAIN 2021.
WBJEE: Placed in the top 0.1% in the state among candidates appearing in WBJEE 2021
Scientific Forum: Selected as a delegate out of 1000+ candidates to represent India at the Asia Pacific Forum for Science Talented 2019.
Case Study: Stood 1st amongst 5000+ participants in the BITS APOGEE, CaseQuesta challenge 2022.

Extracurriculars

Technical Writing: Writer of a series of blogs reviewing papers on ML, DL and AI. [Medium]
NSS Volunteer: Recipient of the gold medal for exceptional service work as an active participant in cleanliness drives, clothes distribution drives and education camps conducted by the NSS in villages near Kharagpur.