Vedant Palit

(+91) 9163389534 | 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

Relevant Coursework: Scalable Data Mining(EX), Big Data Analysis(B), Regression and Time Series Models(EX), Statistical Learning(EX), Transform Calculus(EX), Operations Research(B), Programming and Data Structures(EX), Linear Algebra - Numerical and Complex Analysis(EX)

Publications

- Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP - **Vedant Palit**, Rohan Pandey, Aryaman Arora, Paul Pu Liang ***(ICCV 2023*** *CLVL Workshop).*

- Forgotten Polygons: Multimodal Large Language Models are Shape-Blind - William Rudman*, Michal Golovanevsky*, Amir Bar, **Vedant Palit**, Yann LeCun, Carsten Eickhoff, Ritambhara Singh ***(ACL 2025*** *Findings Track).*

- What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation - Michal Golovanevsky*, William Rudman*, **Vedant Palit**, Ritambhara Singh, Carsten Eickhoff ***(NAACL 2025).***

- WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions - Seyedali Mohammadi, Edward Raff, Jinendra Malekar, **Vedant Palit**, Francis Ferraro, Manas Gaur ***(EMNLP 2024*** *Blackbox NLP Workshop).*

- Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust - Kaushik Roy, Tarun Garg, **Vedant Palit** ***(IEEE CAI 2023).***

- Exploring The Potential of Large Language Models for Assisting with Mental Health Diagnostic Assessments - Kaushik Roy, Harshul Surana, Darssan Eswaramoorthi, Yuxin Zi, **Vedant Palit**, Ritvik Garimella, Amit Sheth ***(ACM Transactions on Computing for Healthcare).***

In Preparation

- One-versus-Others in Vision-Language Models: Steering Model Behavior through Representation Separation - **Vedant Palit**, Michal Golovanevsky, William Rudman, Carsten Eickhoff, Ritambhara Singh *(In preparation for ACL ARR February 2026).*

- Long-term Fairness Dynamics in Performative RL Environments - **Vedant Palit**, Udvas Das, Brahim Driss, Debabrota Basu *(In preparation for ICML 2026; Extended as Master's Thesis Project).*

Theses

- Adaptive Trust Aware Defence in Federated Learning: A Deep Q Network Approach to Detect Weight Poisoning Attacks - **Vedant Palit**, Sayak Roychowdhury (*Bachelor Thesis Project II*).

- Reinforcement Learning Techniques for the Penetration Testing of Computer Networks - **Vedant Palit**, Sayak Roychowdhury (*Bachelor Thesis Project I*).

Research Experience

Research Collaborator | University of Tübingen | Brown University **Providence, RI (Remote)**

Advised by Dr Ritambhara Singh and Dr Carsten Eickhoff *Feb 2024 – Present*

- Analysed the effectiveness of a novel semantic minimal pair and symmetric text replacement corruption scheme, against Gaussian noise-based causal mediation analysis through activation patching and knockouts.
- Evaluated the multi-step math reasoning capability for the LLaVA-1.5, LLaVA-Next, LLaVA-OneVision, Qwen2-VL, Molmo, InternVL, LLaMA-3.2 and Math-PUMA vision-language models.

Research Intern | Inria, University of Lille **Lille, France (Remote)**
Advised by Dr Debobrata Basu *Dec 2023 – Present*

- Proposed and developed a standardised gymnasium environment for the evaluation of algorithmic fairness in the approval of loans through credit worthiness.
- Benchmarked standard RL algorithms such as Linear-Q, Discrete-Q, Proximal Policy Gradient and performative RL algorithms across social welfare, rawlsian, fairness aware and utilitarian reward functions.

Research Collaborator | Carnegie Mellon University **Pittsburgh, PA (Remote)**
Advised by Rohan Pandey, Aryaman Arora and Dr Paul Pu Liang *April 2023 – Dec 2023*

- Proposed and developed a pipeline adapting causal mediation analysis utilising Gaussian Noise injection and activation patching to interpret blackbox architectures of VL transformers.
- Demonstrated the methodology on BLIP and used the COCO-QA dataset to study the internal layer effect of the multimodal text-encoder on model outputs as well as analysed the variation of patching effects with the strength of the noise.

Research Intern | University of Maryland, Baltimore **Baltimore, MD (Remote)**
Advised by Dr Manas Gaur *Nov 2022 – Sep 2024*

- Trained various general and domain-specific models for suicide risk assessment, using the cross entropy and gamblers 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.

Research Collaborator | University of South Carolina **Columbia, SC (Remote)**
Advised by Dr Kaushik Roy *Feb 2023 – May 2023*

- Developed a novel evaluation method to measure error in reconstruction of masked knowledge graph structures from outputs by LLMs.
- Benchmarked models from the GPT Neo family to demonstrate the discontinuity between linguistic fluency and object-level grounding through %Top@5 saturation on higher parameters.

Industry Experience

Machine Learning Intern | JP Morgan Chase **Bangalore, India**
Model Risk Governance and Review Division *May 2025 – July 2025*

- Investigated the internal mechanisms of a LLaMA-3.1 model fine-tuned on confidential JP Morgan data, utilising causal mediation analysis and activation patching.
- Reviewed quantitative models through backtesting and classical interpretability techniques like SHAP and Partial Dependence Plots.

Research Aligned Activities

Conference Reviewing: Reviewed for ACM SIGKDD Workshop 2024, ICLR 2025, NeurIPS Mechanistic Interpretability Workshop 2025, and ICLR 2026.
Mentoring: Guided over 25 students on research directions, career planning, and academic development.
Technical Writing: Authored blogs reviewing papers on ML, DL, and AI. [Medium]
Research Communities: Member of the Association for Computing Machinery (ACM).

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

Adobe AI Challenge: Captained the team of IIT KGP to Gold among 23 competing IITs at Inter IIT Tech Meet 2024.
JEE Advanced: Placed in the top 1.0% 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 to represent India at the Asia Pacific Forum for Science Talented 2019.

Extracurriculars

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.
Blogs on history: Writer of a series of blogs on history and prehistory and their implications on the world. [Blog]
College Societies: An active member of the IIT Kharagpur Quiz Club.