

Parva Mehta

+91-9321565934 | parvamehta26@gmail.com | [linkedin.com/in/parva-mehta](https://www.linkedin.com/in/parva-mehta) | github.com/Parva-26

EDUCATION

Dwarkadas J. Sanghvi College of Engineering

B.Tech in Information Technology

Mumbai, India

Aug. 2024 – May 2028

RESEARCH EXPERIENCE

Research Collaborator

Remote

Mechanistic Interpretability Research Group, In Simplicity Technologies

Jul. 2026 – Present

- Member of a 20+ person international collaboration led by Marko Karbevski studying symmetry-motivated changes to transformer architecture; contributing to the Interpretability/Theory axis.
- Designed and ran causal interpretability experiments (ablations, whole-pathway and query-patching interventions) comparing nonlinear-query transformers against baseline and $W_Q = \text{Id}$ controls across GPT-2 Small, GPT-2 Medium, and GPT-3-Large (2 data-order seeds) using TransformerLens.
- Co-second author of *Beyond Linearity in Attention Projections*, accepted at the SLMs for Agentic Systems Workshop, NeurIPS 2026; currently extending it to a main-track paper and contributing to a second paper on skip connections as gauge fixes (both targeting ICML 2027).

Open-Source Collaborator — CausationEntropy

Remote

Clarkson Center for Complex Systems Science (C3S2), Clarkson University

Sep. 2026 – Present

- Contributing to an open-source Python library for **causal network discovery from multivariate time series** via Optimal Causation Entropy (oCSE), alongside Dr. Kevin Slote (Postdoctoral Researcher, Clarkson University).
- Authored pull requests on information-theoretic estimators (k NN conditional mutual information, Poisson CMI correction) and documentation (under review); library paper planned for JMLR.

Founding Research Member

Remote

Neural Representation Dynamics Lab (NRDL)

Aug. 2026 – Present

- Founding member of an independent mechanistic interpretability group; co-authored an ICLR 2027 submission comparing how absolute (APE) and rotary (RoPE) positional encodings are represented in Vision Transformers, using probing and ablation-based causal analysis.
- Contributing to a new project on **World Action Models** (video world models that jointly predict future frames and robot actions), testing whether hidden states causally encode future-aware cues at inference time.

Research Contributor

Remote

StationHouse

Sep. 2026 – Present

- Contributing to two ongoing LLM research projects: *ConspiracyBench* and *LLM Consciousness Self-Attribution*.

PUBLICATIONS & PREPRINTS

Accepted

- **Beyond Linearity in Attention Projections** [OpenReview]

Jiwon Jeong, Dun Li Chan[†], Andrew Ceniccola[†], Rahul Sharma[†], Simon Sang[†], **Parva Mehta**[†], Pietro Di Nicco[†], Neel Kumar, Karim Saraipour, Marko Karbevski — *SLMs for Agentic Systems Workshop, NeurIPS 2026* — *main-track expansion in preparation for ICML 2027*

Under Review

- **Anchored but Not Protected: When Refusal Erodes During Fine-Tuning, and Whether Holding One Direction Stops It**

Jiwon Jeong*, **Parva Mehta***, Karim Saraipour* — *ICLR 2027*

- **How Vision Transformers Represent Position: A Mechanistic Comparison of Absolute and Rotary Positional Encodings**

Mahmoud Mannes, Nikhil Maturi, **Parva Mehta**, Jiwon Jeong, Aravind Singaram Kannappan — *ICLR 2027*

In Preparation (working titles; author order TBD for all of them)

- **Skip Connections are Gauge Fixes**

Marko Karbevski, **Parva Mehta**, Jiwon Jeong et al. — *Target: ICML 2027*

- **Do World Action Models Use Their Imagined Futures? A Mechanistic Study of Video–Action Coupling**
Mahmoud Mannes, Rajit Rajpal, **Parva Mehta**, Jiwon Jeong et al. (NRDL) — *Target: ICML 2027*
- **Learner-Conditioned Supervision: Model–Data Compatibility in LLM Post-Training**
Jiwon Jeong, **Parva Mehta** — *Target: ICML 2027*
- **Adapt the Teacher or the Student? Tracing What a Crowd-Counting Student Inherits from Its Teacher**
Parva Mehta et al. — *Target: CVPR 2027*
- **Erasers Everywhere? Component Opposition Is Common and Writer-Specific**
Parva Mehta*, Jiwon Jeong* — *Target: ICLR 2027 Workshop*
- **ConspiracyBench**
StationHouse; contributor — *Ongoing*
- **LLM Consciousness Self-Attribution**
StationHouse; contributor — *Ongoing*

*Co-first authors. †Equal contribution (co-second authors).

TECHNICAL SKILLS

Languages: Python, C, C++

ML / DL Frameworks: PyTorch, TensorFlow, HuggingFace Transformers

Interpretability Tooling: TransformerLens, NNsight, SAELens

Core Concepts: Mechanistic Interpretability, Causal Analysis, Causal Discovery, Information-Theoretic Estimation, Transformer Architecture, Large Language Models, Vision Transformers

Scientific Python: NumPy, SciPy, pandas, scikit-learn

Dev Tools: Git, GitHub, Jupyter Notebook, HuggingFace

PROJECTS

IOI Circuit Analysis: Replication + Stability Extension | *PyTorch, TransformerLens, Mech Interp* [GitHub](#)

- Replicated the Indirect Object Identification circuit in GPT-2 small (Wang et al. 2022) using Direct Logit Attribution and path patching to causally verify Name Mover and S-Inhibition heads.
- Extended the original work by probing circuit stability under distractor-token prompt variations; head-importance rankings correlated at $\rho=0.781$ ($p<10^{-30}$) across templates, with 98.2% logit-diff performance retained.

CBAM-ResNet50 on CIFAR-10 | *Python, PyTorch, Attention Mechanisms, Computer Vision* [GitHub](#)

- Integrated **CBAM attention modules** (channel + spatial) into CIFAR-adapted ResNet-50 bottleneck blocks; trained 60 epochs with SGD, cosine annealing and linear warmup.
- Reached **94.65% validation accuracy** vs. 88.49% baseline (+**6.16%**) with only +**0.77%** parameter overhead.

GPT from Scratch | *Python, PyTorch, tiktoken, Mixed Precision Training* [GitHub](#)

- Built a **decoder-only GPT (~30M params)** from scratch in PyTorch (8 layers, causal multi-head attention) without HuggingFace abstractions.
- Trained on an **~800M-token corpus** for ~53K steps with AMP mixed precision and gradient accumulation on free-tier T4 GPUs, reaching validation loss 4.63 (perplexity ~103).

EXTRACURRICULAR ACTIVITIES

Research Lead

DJ InIT.ai

Mumbai, India

Aug. 2026 – Present

- Runs DJSCE's student machine-learning research club: paper-reading sessions, technical talks, Kaggle contests, and group discussions.
- Coordinates outside research collaborations for members and mentors juniors toward writing their own research papers.