

VISHAL PRAMANIK

Ph.D. Researcher | Explainable AI, LLM Safety & Multi-Agent Systems

Gainesville, FL | +1 (786) 553-7590 | vishalpramanik@ufl.edu

[Google Scholar](#) | [LinkedIn](#) | [Research Portfolio](#)

RESEARCH PROFILE

Ph.D. student at the University of Florida seeking a Summer 2027 research internship in LLM/VLM interpretability, trustworthy AI, or multi-agent collaboration. Research connects explanations of model behavior with interventions for attribution, jailbreak defense, machine unlearning, and reliable agent coordination. First-author work at NeurIPS, ICLR, CVPR, and EMNLP, including a CVPR Highlight; two years of firmware engineering experience at Intel.

EDUCATION

University of Florida | Ph.D., Computer and Information Sciences | 2026 - Present

Advisor: Prof. Sumit Kumar Jha

Florida International University | Doctoral study | 2024 - 2025

Computer and Information Sciences; GPA: 4.0/4.0. Transferred to UF.

Indian Institute of Technology Bombay | M.Tech., CSE | 2020 - 2022

GPA: 9.01/10; Certificate of Excellence in Research.

West Bengal University of Technology | B.Tech., CSE | 2015 - 2019

GPA: 8.9/10.

RESEARCH & INDUSTRY EXPERIENCE

Graduate Research Assistant | **University of Florida** | 2026 - Present

- Develop training-free and intervention-based methods to explain and control language-model behavior, including token attribution, adversarial robustness, and machine unlearning.
- Investigate distributed harmful intent, decentralized coordination, and task, role, and state drift in multi-agent systems using both model-internal and black-box approaches.
- Build PyTorch and Transformers evaluation pipelines with ablations, sensitivity studies, and statistical analyses across language and vision tasks.

Silicon Firmware Development Engineer | **Intel** | July 2022 - July 2024

- Developed DDR5 memory-initialization features, including unregistered-DIMM support and reference-voltage tables for server platforms.
- Contributed to boot flow, error handling, memory mapping, thermal control, and performance optimization; worked with validation engineers on stress testing and failure analysis.

Master's Researcher | **IIT Bombay** | 2020 - 2022

- Conducted transformer-based NLP research in language generation, sentiment analysis, information extraction, translation, and automatic essay grading.
- Contributed to a movie-script generation assistant developed for ErosNow.

Summer Trainee | **Centre for Development of Advanced Computing** | July - August 2018

- Developed a user interface for monitoring and regulating server temperatures in a data-center environment.

TECHNICAL SKILLS

Languages: Python, C, C++, Java, SQL; Git and LaTeX.

Frameworks: PyTorch, TensorFlow, Transformers, scikit-learn, NLTK, OpenCV.

Analysis: NumPy, Pandas, Seaborn; ablation design and statistical evaluation.

Methods: Mechanistic interpretability, attribution, LLM safety, multi-agent coordination, machine unlearning, NLP, computer vision, reinforcement learning.

VISHAL PRAMANIK | RESEARCH PUBLICATIONS

Selected papers most relevant to research internships in interpretability, model safety, and agentic AI. First-author contributions are identified below.

INTERPRETABILITY, MODEL CONTROL & UNLEARNING

Hessian-Enhanced Token Attribution (HETA): Interpreting Autoregressive LLMs.

ICLR 2026 | First author

Hessian-enhanced attribution for interpreting autoregressive language-model predictions; connects model-internal analysis with token-level explanations.

Jailbreaking the Matrix: Nullspace Steering for Controlled Model Subversion.

ICLR 2026 | First author

Studies controlled model subversion through nullspace interventions, linking internal representation analysis with language-model security evaluation.

Selective Amnesia using Contrastive Subnet Erasure for Class Level Unlearning in Vision Models.

CVPR 2026 | Highlight | First author

Investigates selective class-level forgetting through contrastive subnet erasure in vision models, connecting targeted model modification with machine unlearning.

NSF-CoT: Neuro-Symbolic Formal Verification of Chain-of-Thought Faithfulness in Contextual Question Answering.

Findings of ACL 2026 | First author

Combines neuro-symbolic reasoning and formal verification to examine chain-of-thought faithfulness in contextual question answering.

MULTI-AGENT SAFETY & RELIABLE COORDINATION

Harmless in Pieces, Harmful in Motion: Detecting Multi-Agent Jailbreaks.

NeurIPS 2026 | Accepted poster | First author

CITADEL tracks collective movement toward harmful objectives across agents, tools, and memory to detect distributed jailbreak behavior.

Unified Resource-Grounded Coordination Protocol for Orchestrator-Free Heterogeneous Multi-Agent Systems.

NeurIPS 2026 | Accepted poster | First author

URCP coordinates heterogeneous agents without an external orchestrator using a common supermajority decision rule and bounded repair.

Drift No More: Spectral Locking for Long-Horizon Open-Weight Multi-Agent AI.

EMNLP 2026 | Accepted main conference | First author

ATLAS intervenes on hidden representations to jointly address task drift, role drift, and state staleness during long-running agent collaboration.

ATC: Agent Trajectory Control for Role-Contract Enforcement in Black-Box Multi-Agent Systems.

EMNLP 2026 | Oral | Coauthor

Enforces agent role contracts through output scoring and constrained regeneration, supporting role control when model internals are inaccessible.

RECOGNITION & RESEARCH SERVICE

CVPR 2026: Highlight paper selection for class-level machine unlearning.

IIT Bombay: Certificate of Excellence in Research for master's thesis research.

IEEE NAECON 2024: Second Place Poster Award, Human Fatality Estimation in Aircraft Accidents.

Peer review: ICLR, NeurIPS, AAAI, and CVPR.