Sumit Yadav

Mechanistic Interpretability | AI Alignment & Safety | NLP

076bct088.sumit@pcampus.edu.np | +977-9819856148 | tatva.sumityadav.com.np | Google Scholar | github.com/rockerritesh | LinkedIn

Research Statement

I work at the intersection of **mechanistic interpretability, AI alignment, and robustness**. My research addresses how safety-aligned LLMs fail silently — through over-refusals, geometric misrepresentation, and surface-level triggers — and how the internal structure of representations can be understood and steered to make models safer and more reliable. I combine representation-space analysis, adversarial probing, and systems engineering to develop empirically grounded, technically strong methods. My work has been accepted at **ACL 2026 (CORE A*)**, and I am especially interested in extending interpretability and safety to multilingual and low-resource settings.

Selected Research & Publications

SafeConstellations: Mitigating Over-Refusals in LLMs Through Task-Aware Representation Steering | *ACL 2026 Main Conference (CORE A*)* 2026

- **Problem:** Safety-aligned LLMs persistently refuse inputs containing harmful content even when reframed with benign tasks (e.g., sentiment analysis, language translation), degrading utility in production applications without improving security.
- **Discovery:** LLMs follow distinct *"constellation"* trajectory patterns in embedding space as representations traverse layers — each NLP task maintains consistent trajectories that shift predictably between refusal and non-refusal states, enabling mechanistic interpretability of over-refusal behavior.
- **Method:** Introduced **SafeConstellations**, an inference-time *task-aware trajectory-shifting* approach that tracks task-specific patterns and selectively guides representations toward non-refusal pathways *only* on tasks prone to over-refusal — no retraining required.
- **Result:** Reduces over-refusal rates by up to **73%** with minimal impact on utility — a principled, conditional approach to scalable safety alignment. *Research Area: Safety and Alignment in LLMs* (ACL 2026, Submission #5757).

On the Relationship Between Representation Geometry and Generalization in Deep Neural Networks | *Pre-Print 2026* 2026

- Established that **effective dimension** — an unsupervised geometric metric — strongly predicts generalization in both vision and language models (partial $r = 0.75$ across 52 ImageNet classifiers, 13 architectures).
- Cross-domain results hold for NLP encoders and decoder-only LLMs, offering a label-free diagnostic applicable to safety-relevant distributional shift detection and **robustness evaluation**.

Geometric Phases of Mechanism Formation in Neural Networks | *Working paper* 2026

- Traces how internal mechanisms form across training using linear probes, CKA, and targeted ablations (CIFAR-10/100) — linking representation geometry to when and how capabilities emerge.

Can maiBERT Speak for Maithili? | *LoResLM @ ACL 2026* 2026

- Built the first monolingual BERT for Maithili (spoken by ∼50M people); **87.02% accuracy** on news classification, outperforming multilingual baselines (MuRIL, NepBERTa).
- Demonstrates principled approaches to low-resource, underrepresented communities — relevant to **fairness, equity, and AI access** dimensions of safety.

Revolutionizing Currency Security with YOLOv8 | *J. Bus. Econ. Stud., 2024* 2024

- Applied YOLOv8 to counterfeit Nepali banknote detection; achieved **True Positive Recall of 0.986** — a concrete example of high-severity misuse (fraud) mitigation via CV.

Professional Experience

Astha.ai USA (Remote)
AI Researcher — Safety & Agentic Systems *May 2025 – Present*

- **Zero-Trust Agentic Oversight:** Designed a Zero-Trust framework for Autonomous Agents in which every tool call, memory retrieval, and inter-agent message carries a cryptographic identity verified against a strict policy engine — directly addressing **agentic oversight** at the architectural level.
- **MCP-Proxy (Hexagonal Architecture):** Built the core security proxy for Model Context Protocol servers using the Ports & Adapters pattern, decoupling security policy logic from transport — enabling independent audit of safety rules without live network dependencies.
- **Policy Engine & RBAC:** Engineered a two-tier policy engine (v1.0: allow/deny lists; v2.0: conditional logic) with Role-Based Access Control (Admin/User/Guest), providing **granular, auditable permission management** for LLM tool usage.
- **MCP-Scanner — Vulnerability Detection Platform:**
  - Integrated **78+ attack techniques** mapped to the MITRE ATT&CK framework for systematic adversarial evaluation of AI agent infrastructure.
  - Leveraged the Claude API to intelligently fuzz and enumerate vulnerabilities in deployed MCP servers.
  - Built *SAFE-T1102* (Prompt Injection Defense) and *SAFE-M-1* (Tool Poisoning Defense) — dual empirical **safety evaluation** modules.

AMNIL Technologies Nepal (Hybrid)
AI Engineer — RAG & Infrastructure *Jun 2024 – May 2025*

- **Guardrails & Evaluation:** Integrated NeMo Guardrails to prevent hallucination and enforce topical boundaries; built an **LLM-as-a-Judge** evaluation framework for benchmarking response accuracy — an early instance of scalable safety evaluation pipelines.
- **Self-Hosted LLM Infrastructure:** Deployed and optimized Llama 3, Mistral, and Qwen via vLLM; reduced latency by **40%** while maintaining output quality, enabling cost-effective safety-evaluation at scale.
- **Advanced RAG with Hybrid Search:** Implemented Sparse + Dense hybrid retrieval in Qdrant, improving retrieval accuracy by **35%** on multi-hop domain-specific queries.

GradeUp Educations Kathmandu, Nepal
Data Team Lead *Jan 2022 – Jun 2024*

- Led the data team building learning agents/chatbots for students, an automated **grade-evaluation** system, and semantic-similarity matching between questions and answers (NLP + image processing).

DeepLearning.AI Global (Remote)
GAN Specialization Mentor *Aug 2021 – Present*

- Technical mentor for the **GANs Specialization**, supporting hundreds of students globally on loss functions (Minimax, Wasserstein), training stability, and generative model robustness.

Key Safety & Alignment Projects

SAFE-MCP Security Framework | *Open Source — github.com/rockerritesh* 2024–Present

- Core maintainer of the **Security Analysis Framework for Evaluation of Model Context Protocol (SAFE-MCP)**: a systematic adversarial evaluation suite for LLM agent infrastructure.
- Authored detection rules and proof-of-concept exploits covering: *SAFE-T1601* (MCP Server Enumeration), *SAFE-T1703* (Tool-Chaining Pivot attacks), *SAFE-T1110* (Multimodal Prompt Injection via steganography), and *SAFE-T1604* (Server Version Enumeration).
- Provides empirical, reproducible **safety evaluations** for a class of agentic systems that are increasingly deployed but understudied from a security and alignment perspective.

Telebot-Claude-Bridge | *Open Source — GitHub* 2025–2026

- Production-grade bridge enabling remote, **permission-gated** control of Claude Code CLI via Telegram — demonstrating human-in-the-loop oversight for autonomous coding agents.
- Key safety-relevant features: inline Allow/Deny/Allow-All approval for every tool call, multi-session audit logging, and ESC interrupt support — a practical instantiation of **interruptibility and corrigibility** in agentic systems.

NPL Coders — National AI Safety & Data Science Community Sep 2023–Present

- Founded and lead Nepal's largest data science competition platform on Kaggle and HackerRank; organized internet safety education workshops, expanding the regional AI safety talent pipeline.

Technical Proficiency

Interpretability & Alignment: Representation steering, mechanistic interpretability, over-refusal mitigation, prompt injection defense, tool poisoning detection, RLHF, guardrails, LLM-as-a-Judge evaluation.
Adversarial & Security Evaluation: MITRE ATT&CK mapping, protocol fuzzing, SSRF/injection detection, multimodal prompt injection, 78+ agentic attack techniques (SAFE-MCP).
LLMs & NLP: Large Language Models (Llama 3, Claude, GPT-4), Fine-tuning (PEFT/LoRA), RAG Pipelines, Multilingual/Low-Resource NLP, Tokenizer training, Embedding alignment.
Agentic Systems: Model Context Protocol (MCP), Multi-Agent Frameworks, LangChain, LlamaIndex, Semantic Routing, tmux/session orchestration, Telegram Bot API.
Infrastructure & Tools: Python (Expert), PyTorch, TensorFlow, vLLM, FastAPI, Docker, Kubernetes, Vector DBs (Qdrant, ChromaDB, Pinecone), GitHub Actions.
Research Methods: Representation geometry, information-theoretic metrics, empirical NLP evaluation, adversarial benchmarking, statistical analysis.

Education

Pulchowk Engineering College, IOE — Tribhuvan University Kathmandu, Nepal
Bachelor of Computer Engineering *2019 – 2024*

- **Major Project:** "Evaluating Auto-Encoder Transformer Language Models for Maithili Text Classification" — the empirical foundation for maiBERT and subsequent ACL 2026 publication.
- **Relevant Coursework:** Artificial Intelligence, Distributed Systems, Network Security, Big Data Technologies, Compiler Design.

Honors & Awards

Winner, GritFeat AI Hackathon (2023): *SWIFT* — wearable LSTM fall-detection for the elderly (79.86%).
1st Runner-Up, Locus Dataverse (2023): NLP classification of imbalanced research-paper abstracts.
1st Runner-Up, Docsumo DataRush (2022): Abstract classification into 158 classes (SVC + TF-IDF).
Best AI Project, DELTA 3.0 (2023): *Nepali Harvest* — crop-disease prediction & harvest timing.
Winner, IT-Meet Image Challenge (2022): CV classification of Nepali ballot-paper images.
Winner, LogPoint Capture The Flag (2022): Binary exploitation & forensics.