Swapnil **Parekh**

AI Scientist — Mechanistic Interpretability, Adversarial Robustness, Agentic Systems | Open Source @vLLM | Kaggle Expert

swapnilbp100@gmail.com • +1 (631) 703 7752 • Google Scholar • LinkedIn • GitHub

RESEARCH FOCUS

Mechanistic Interpretability
Adversarial Robustness & Safety
Agentic LLM Systems
RL Alignment (RLVR / GRPO)
Explainable AI
VLM / Multimodal Safety

EDUCATION

New York University
MS Computer Engineering
| 2021 – 2023

VJTI Mumbai
BS Computer Science
| 2017 – 2021

London School of Economics
ML Summer School | Topped-A+

TECHNICAL SKILLS

ML / DL: PyTorch, TensorFlow, vLLM, JAX, HuggingFace

Agents: LangGraph, MCP, RAG, Tool-use / Function Calling

Infra: k8s, OpenShift, Docker

Core: Python, SQL, Git, Ray

AWARDS

Kaggle Competition Expert
Jigsaw Toxicity (Top 9%)
Google QA Labelling (Top 8%)

NVIDIA Global AI Challenge
2nd Place — DL CUDA/TensorRT

SoftBank Forex Prediction
12th / 2500 teams

IBM AI Hackathon, IIT Bombay
1st Place

Google Certified
TensorFlow Developer

OPEN SOURCE

vLLM — Contributed prefix-adapter support for fine-tuned model serving (merged)

RESEARCH

80+ citations

[1] **S. Parekh**. **CIRCUS**: Circuit Consensus under Uncertainty via Stability Ensembles. *arXiv*, 2026. — *Uncertainty-aware circuit discovery; ∼40× smaller consensus circuits; validated on Gemma-2 & Llama-3.2.* [Link]

[2] **S. Parekh**. Thinking Wrong in Silence: Backdoor Attacks on Continuous Latent Reasoning *arXiv*, 2026. — *Fundamentally new attack surface for latent reasoning models* [Link]

[3] **S. Parekh**. **CaptionFool**: Universal Image Captioning Model Attacks. *IntelliSys (Springer)*, 2026. — *94–96% targeted success modifying only 1.2% of image patches; evades content moderation.* [Link]

[4] **S. Parekh**, Y.S. Kumar, C. Chen. **MINIMAL**: Mining Models for Universal Adversarial Triggers. *AAAI*, 2022. — *Data-free UATs matching data-dependent baselines.* [Link]

[5] **S. Parekh**, Y.S. Kumar, J.J. Li, C. Chen. AES Systems Are Both Overstable And Oversensitive: Evals & Defenses. *Dialogue & Discourse*, 2023. [Link]

[6] **S. Parekh**, P. Shukla, D. Shah. Attacking Compressed Vision Transformers. *FICC (Springer)*, 2023. [Link]

[7] D. Mahata, D. Gautam, **S. Parekh** et al. **LDKP**: Keyphrases from Long Scientific Documents. *CIKM*, 2022. | Adobe Research. [Link]

EXPERIENCE

Intuit | Senior AI Scientist, Expert AI Science
Oct 2024 – Present | Mountain View, CA

- **Intuit Assist (Patented):** Architected secure multi-agentic LLM system with RAG + fine-tuned SLM tool calling for tax QnA and automated resolution workflows across TurboTax, QuickBooks, and Credit Karma.
- Designed and shipped the **first MCP (Model Context Protocol) implementation at Intuit** — now core to the **Intuit × Anthropic partnership**. Serves **100 M customers**; reduced customer serving time by **30%** and increased response rate by **40%**.
- **Agentic Upsell System:** Shipped product recommendation engine using **deep-learning twin-tower architecture** (multimillion-dollar captured in upsell conversions). Built sales pitching agent with constrained-decoding guardrails and **GRPO rollouts + RLVR** with hard/soft reward signals on foundation model.

IBM | AI Scientist, Chief Analytics Office
June 2023 – Oct 2024 | New York, NY

- Core team, **watsonx.ai LLM Inference engine** (open source **@vLLM**, now a PyTorch Foundation project used by thousands of production deployments). Contributed **prefix-adapter serving support** enabling efficient multi-tenant fine-tuned model inference, plus latency and throughput optimizations.
- Led delivery of **PLG SaaS churn prediction** using stacked gradient boosting with **SHAP + LLM-generated explainability** narratives for non-technical stakeholders; multimillion-dollar contracts retained per year.

AI Science Intern — June 2022 – Aug 2022 | New York, NY

- Built skill recommendation engine with **self-supervised Siamese Networks** (triplet loss) for improved representation learning on sparse internal HR data; deployed to reduce hiring costs.

Also at **Orenda Power** (stochastic optimization for energy markets) • **Spherical Defence Labs** (Transformer network request anomaly detection, SOTA) • **Tata Digital** (Ecommerce AI) • **MetLife** (Applied ML)