Sripad Karne

858-275-3303 | sk5695@columbia.edu | linkedin.com/in/sripad-karne | Google Scholar | US Citizen

Education

Columbia University — New York City, NY
Master of Science (**M.S.**) in Data Science — *Expected Dec. 2027*

University of California, San Diego — San Diego, CA
Bachelor of Science (**B.S.**) in Cognitive Science w/ specialization in Machine Learning — *Sep. 2021 – June 2025*

Publications

How Far Do Auto-Interpretation Labels Generalize? A Controlled Study Across Languages, Scripts, and Rewordings
Submitted to EMNLP 2026 (main track, via ARR); preprint on arXiv

- **Solo author**. Factorial design isolating script, language, wording, and meaning using Serbian digraphia, with English and Russian controls, to test whether SAE auto-interpretation labels generalize beyond English.
- Finds that content-labeled SAE features miss the same meaning in Serbian up to 4x more often than within English, with Cyrillic missed up to 1.35x more than Latin despite the two being deterministic transliterations; the gap widens with network depth and is invisible from the labels themselves.

One Language, Two Scripts: Probing Script-Invariance in LLM Concept Representations
ICLR 2026 UCRL Workshop

- **Solo author**. Presented at ICLR 2026 in Rio de Janeiro, Brazil. Uses Serbian digraphia as a controlled testbed to test whether SAE features capture abstract semantics or surface-level token patterns across the Gemma family (270M–27B).
- Compares SAE feature activations on parallel Latin and Cyrillic renderings of identical text, isolating script from semantics under fully disjoint tokenizations.

Silent Multilingual Failure in Production Safety Probes: A Controlled Cross-Language Audit
In preparation; to be submitted to BlackboxNLP 2026 (EMNLP Workshop)

- **Solo author**. Audits the production linear-probe recipe used in deployed safety monitors at frontier labs for silent multilingual failure, measuring recall/FNR at the frozen English-calibrated threshold across English, Serbian (Latin and Cyrillic), Russian, and Telugu.
- Uses Serbian digraphia for script-controlled causal attribution, isolating tokenization and script coverage from language identity and semantics

Client-Side Optimization for LLM Agent Pipelines: Combinatorial Model Selection with Sample-Efficient Search
Submitted to NeurIPS 2026 (main track)

- **Second author**. Joint work with Microsoft Research, Cornell, and Columbia. Bandit-based approach to model selection in multi-step agent pipelines.

AgentOpt
Open-source library and technical report, arXiv 2026

- **Second author**. Framework-agnostic optimization library for client-side model selection in multi-agent LLM systems; released as an installable open-source package with a technical report documenting system design, API, and usage.

Experience

Independent Researcher — Nov. 2025 – Present
Mechanistic Interpretability — *New York City, NY*

- Designed and executed a **solo** research program on **cross-lingual SAE interpretability** using Serbian digraphia as a controlled testbed, producing three solo-authored papers: an accepted **ICLR 2026 workshop** paper, a main-track **EMNLP 2026** submission, and a **BlackboxNLP 2026** probe-generalization audit in preparation
- Built the full experimental stack end to end: a factorial multilingual evaluation suite on FLORES+ with native-speaker validation, activation extraction and SAE analysis via **TransformerLens** and **sae-lens** across the **Gemma** and **Llama** families (270M–27B), and auto-interpretation label evaluation against Neuronpedia's deployed pipeline

- Built a meaning-matched, **native-speaker-validated harmful-content probing dataset** spanning English, Serbian (Latin and Cyrillic), Russian, and Telugu, and reimplemented the production linear-probe recipe (mean-pooled, frozen-threshold) with **layer and scale sweeps** across the Gemma family (270M–27B), measuring recall/FNR at the deployed English-calibrated operating point
- Currently extending the cross-lingual evaluation paradigm to newer interpretability methods, including **Natural Language Autoencoders**, testing whether activation-to-language methods generalize across languages and scripts; targeting **ICLR 2027**

AI Engineer Intern May 2026 – Present
IBM *San Francisco, CA*

- Design and deploy production LLM systems for enterprise clients, building agentic and RAG pipelines on watsonx and open-source model stacks (vLLM-served Llama/Granite), including retrieval evaluation and latency/cost optimization

Graduate AI Researcher Jan. 2026 – May 2026
Data, Agents, and Processes (DAP) Lab (advised by Prof. **Tianyi Peng**) *New York City, NY*

- Lead development of AgentOpt, a framework-agnostic optimization layer for multi-agent LLM systems that assigns heterogeneous models across agents to optimize accuracy, latency, and cost, in collaboration with researchers at **Columbia**, **Cornell**, and **Microsoft Research**
- Ran end-to-end benchmarking across **4 agentic benchmarks** (GPQA, BFCL, HotpotQA, MathQA), multiple model families, and agent frameworks (**CrewAI**, **LangGraph**, **LlamaIndex**); contributed to the open-source package
- Co-authored **AgentOpt** technical report (arXiv 2026), presenting the first framework-agnostic package for client-side model selection in agentic workflows; manuscript submitted for **NeurIPS 2026** main track on sample-efficient search algorithms for cost-aware model selection

Awards

NVIDIA & Dell GB10 Agentic AI Hackathon – Winner (Agentic AI Track)

- Selected as **1 of 40 participants from 600+ applicants** and awarded **1st place** for building **Guardian AI**, a real-time, GPU-accelerated agentic safety intelligence system on the NVIDIA **Grace Blackwell** platform
- Designed and implemented a **multi-agent reasoning framework** for safety-critical video understanding, with temporal event localization and evidence-backed risk assessment over **live video streams**, applying **human-in-the-loop oversight** to ensure model decisions remained interpretable and auditable
- Optimized end-to-end vision and inference pipelines for low-latency **GPU-accelerated** hazard detection, with explicit visual grounding of model outputs to support transparent failure analysis and reduce risks from opaque autonomous decision-making

Technical Skills

Languages: Python, C++, SQL, R
ML & Research: PyTorch, TransformerLens, sae-lens, HuggingFace Transformers, CUDA, Distributed Training (DDP), Multi-GPU Optimization
Infrastructure: AWS, GCP, Linux, Git, LangChain, LlamaIndex, CrewAI
Relevant Coursework: High Performance Machine Learning, Neural Networks and Deep Learning, Advanced Machine Learning Methods, AI Engineering, Artificial Intelligence of Things, Algorithms for Data Science, Modeling and Data Analysis