Akshit Kumar

+91-9622690011 | akshit.kumar@research.iiit.ac.in | linkedin.com/in/akshit-kumar-229a06147 |
github.com/komikat

Research Interests

Multilingual generalization in large language models and how such representations emerge, evolve, and transfer across languages. Mechanistic interpretation of these processes to improve model safety, alignment, and reliability for real-world deployment. Additional focus areas: alignment of autonomous LLM agents, rigorous evaluations and benchmarks, robustness and red-teaming.

Education

International Institute of Information Technology Hyderabad, IN
B.Tech in Computer Science and MS (by Research) in Computational Linguistics *Oct. 2021 – Present*

Experience

ML Intern Feb 2025 – Present
Deccan.ai *Hyderabad*

- Built human-in-the-loop evaluation service for agent benchmarking; used in the Anthar Study on 6 coding agents over 43 tasks.
- Designed an agentic testbed with synthetic tools and databases to stress-test multi-turn LLM agents and log failure modes.
- Analyzed RAG pipelines on complex queries; identified retrieval and generation errors and improved retrieval accuracy.
- Shipped a FastAPI backend to serve agent interactions to a web UI that enables efficient annotator validation.

Undergraduate Researcher Sep 2023 – Present
Language Technologies Research Center *IIIT Hyderabad*

- Built multilingual dataset pipelines and a dependency parser; evaluated with UAS/LAS and error analyses.
- Fine-tuned cross-lingual models (XLM-R, custom self-attention) for low-resource Indian languages.
- Investigating the emergence of multilingual abstractions in LMs using probing and causal interventions.

GSoC Participant 2024
Haskell Language Server *Remote*

- Improved test-suite performance and stability; optimized CI/CD workflows across a large open-source codebase.

Publications

Do Multilingual Transformers Encode Paninian Grammatical Relations? A Layer-wise Probing Study (2025) — A. Kumar, D. Sharma, P. Krishnamurthy. *Depling 2025*.

Probed XLM-R, mBERT, and IndicBERT across seven languages; found syntactic peaks in middle layers and lexical features earlier.

Projects

Adversarial Robustness of Vision–Language Models — tested safety bypasses in Llama Guard 3 Vision with adversarial perturbations; built frequency-based defenses; evaluated with PGD and Grad-CAM.

Grounding Small LMs on Structured Data — improved grounding and factuality via multi-hop RAG over knowledge graphs; teacher–student distillation yielded +15% accuracy over baseline.

Domain-Adaptive Text Generation Detection — designed black-box detectors robust to domain shift; achieved 92% F1 on held-out data.

Technical Skills

Evaluation: dataset design, LLM-as-judge, HITL pipelines, RAG evals

ML/NLP: PyTorch, Hugging Face, LangChain, SFT/RLHF, spaCy, NLTK

Data: Polars, pandas, NumPy; multilingual preprocessing

Systems: FastAPI, Docker, Linux, AWS; Emacs; MCP, Ollama

Research: alignment, mechanistic interpretability, robustness, multilingual NLP

Open Source

Haskell Language Server — contributed performance optimizations and improved test coverage during GSoC 2024.