

Abhinav Rajeev Kumar

abhinavrkumar.com [@abku2504@gmail.com](mailto:abku2504@gmail.com) [in/abhinav-rajeev-kumar](https://in.linkedin.com/in/abhinav-rajeev-kumar) github.com/MajorTimberWolf [Google Scholar](https://scholar.google.com/citations?user=abhinavrkumar)

Education

May 2026 Aug 2022	SRM Institute of Science & Technology <i>B.Tech in Computer Science and Engineering</i> CGPA: 9.50/10.0	Chennai, India
----------------------	--	----------------

Experience

Present June 2025	Lossfunk  <i>Researcher</i> Researching LLM/agent alignment and AI4Science. Studying source-authority bias and autoresearch failure modes, especially task-driven attractor spaces that are difficult to steer out of.	Bangalore, India
Oct 2025 May 2025	Darko Marinov's Research Group, UIUC  <i>Undergraduate Researcher Advisor: Prof. Darko Marinov</i> Engineered LLM workflows for semantic evaluation on software mutants, classifying code changes as equivalent. Developed analysis scripts to identify “fishy” mutants by extending BugFarm ICSE heuristics.	Remote
May 2024 May 2023	Google (Gemini Team) <i>Software Engineer Intern</i> Refined inappropriate question detection with XAI/LIME (20% FP reduction) and applied chain-of-thought prompting for multi-step reasoning (24% F1 improvement).	Remote

Publications

[* = Equal Contribution]

[2026]	Trust Me, I'm Absolutely Right: When LLM Truth-Seeking is Actually Self-Confirmation [PDF] Abhinav Rajeev Kumar, Paras Chopra <i>6th Workshop on Trustworthy NLP (TrustNLP), ACL 2026 [Oral]</i>	[ACL-W]
[2026]	METIS: Mentoring Engine for Thoughtful Inquiry & Solutions [PDF] Abhinav Rajeev Kumar, Dhruv Trehan, Paras Chopra <i>AI For Scientific Research Workshop, AAAI 2026</i>	[AAAI-W]
[2025]	Analyzing Quantization in Multilingual Transformer Models for Indic Translation [PDF] Praneeth Kasiraju*, Abhinav Rajeev Kumar*, Ramprasath M, Manikandan N <i>IEEE Conference on Computational Methods for Sound Science (CMSS) 2025</i>	[IEEE]

Select Projects

Urgency Steering: Do Models Know When to Stop? <i>Research Project Technical Writeup</i> ► Tested whether LLMs act on deadlines. Qwen and Gemma register the budget and their internals predict completion, but steering toward finishing only made ending tokens cheaper, not the decision smarter.	Aug '26
Personality-Disorder: Circuits vs Vectors <i>MATS Application Technical Writeup</i> ► Asked whether CAA steering vectors act through the circuits producing sycophancy in Llama-3.1-8B-Instruct. Path patching localized no circuit and extraction-based steering backfired, pointing to a distributed behaviour.	Sep '25
Joker-LLM <i>Research Project Technical Writeup Source Code </i> ► Implemented PlanSearch for generating “funny” jokes by using multi-dimensional evaluations and bias mitigations via LLM-as-a-judge methodology.	May '25 – Jun '25
RAO: Recursive Agent Orchestrator <i>Research Project Technical Writeup Source Code </i> ► Developed an agentic system that spawns and manages LLM-driven agents via a master-delegator model in the form of layers, built for complex multi-agent workflows.	Jan '25 – Mar '25