

Suresh Raghu

New Delhi, India | +91-7904210795 | sureshraghu0706@gmail.com

 Website  Scholar  GitHub  LinkedIn  arXiv

RESEARCH INTERESTS

Long-horizon LLM agents · RL beyond verifiable rewards · Evaluation & UQ · Open-ended problem solving

EDUCATION

Indian Institute of Technology Madras

B.Sc. in Programming & Data Science

Tamil Nadu, India

Jan. 2021 – May. 2026

Vellore Institute of Technology Bhopal

B.Tech. in Computer Science and Engineering

Madhya Pradesh, India

Oct. 2020 – Jun. 2024

PUBLICATIONS

- [1] **Proper Scoring Rules for Agentic Uncertainty Quantification.** [Paper] [Poster]
Suresh Raghu*, Satwik Pandey*, Shashwat Pandey [NeurIPS '26, ICML-W '26]
- [2] **SELFDUBT: Uncertainty Quantification for Reasoning LLMs via the Hedge-to-Verify Ratio.** [Paper] [Code] [Poster]
Satwik Pandey*, Suresh Raghu*, Shashwat Pandey [ICML-W '26]
- [3] **Don't Blink: Evidence Collapse during Multimodal Reasoning.** [Paper] [Code] [Poster]
Suresh Raghu*, Satwik Pandey* [COLM-W '26]
- [4] **Confidently Wrong, Silently So: Auditing Undetectable Failures of a Deployed On-Device Language Model.** [Paper]
Shashwat Pandey, Satwik Pandey, Suresh Raghu [NeurIPS-W '26]

RESEARCH & ENGINEERING EXPERIENCE

Looni Lab, Carnegie Mellon University

Research Collaborator

Remote

Jul. 2026 – Present

Advisors: Prof. Niloofar Mireshghallah, Prof. Aviral Kumar

- Work toward training language-model agents for molecular discovery with supervised fine-tuning and RL, focusing on iterative experimentation, scientific tool use, and learning from computational feedback.
- Develop containerized molecular-design environments with verifiable rewards and trajectory logging to support agent training and evaluation on SMDD-Bench.
- Investigate how agent representations, persistent memory, and evaluation budgets influence multi-step scientific search, informing the design of training tasks and feedback.

VFS Global

Senior Manager, AI (since Aug. 2026)

New Delhi, India

May 2024 – Present

- Founding lead of AI engineering (team of 8+), building LLM and VLM systems for document understanding and fraud detection across **30K+ visa applications per day** in multiple countries. Our document-processing platform won the **Aegis Graham Bell Award** (2026, Innovation in Gen AI).
- Built a case-review agent with a sandboxed code interpreter that cross-checks 25+ documents per application, runs calculations and conversions across them, and writes a review report flagging gaps, cutting per-case review time **6x (30+ to under 5 minutes)**.
- Built uncertainty-gated document extraction: a token-logit signal triggers semantic-entropy estimation on suspect fields, and low-confidence outputs escalate to a larger model or human review. Cut critical extraction errors by **87%**.

- Trained Qwen2.5-VL-7B online on streaming passport data (never stored, for PII compliance) with weekly checkpoint gating; it **matched Gemini 2.5 Pro** across **110+ nationalities** and replaced it in production.

Adani Digital Labs

Applied Machine Learning Intern

Gurugram, India

Jun. 2023 – Apr. 2024

- Built a few-shot image classifier from frozen DINOv2 embeddings and an SVM, trained with limited labels; reached **93% precision** on 500+ held-out images and deployed it to auto-tag untagged hotel imagery across **~150K properties**.
- Built multimodal product search over Adani’s airport retail inventory (**~20K SKUs**), using CLIP for text and image retrieval and GPT-4 Vision for attribute extraction.

SELECTED RESEARCH PROJECTS

Harness Engineering for Scientific Agents (with K. Han, A. Kumar, N. Mireshghallah)

[Blog]

- Asked whether changing only the harness (tools, state, observations) can fix scientific-agent failures; diagnosed failure modes of frontier models and Qwen3.5-9B on SMDD-Bench across lead optimization, interaction-point discovery, and scaffold hopping.
- Redesigned harnesses around those diagnoses (an authoritative candidate ledger with cached oracle calls; ligand-side probe molecules), improving lead optimization from **13 to 33/68 tasks** and interaction-point recovery from **10 to 30/75** without retraining.
- Compared manual design against automated harness search (Meta-Harness, **2,640 trajectories**): manual design won where gains needed new representations or evidence (lead optimization, $p = 5 \times 10^{-4}$), while automated search found the best scaffold-hopping harness ($2 \rightarrow 5/52$).

Proper Scoring Rules for Agentic Uncertainty (NeurIPS [1])

[Paper]

- Asked whether an agent’s per-step uncertainty means what it says, not just whether it ranks well; showed that AUROC, Trajectory ECE (resolution-blind), and scalarized Trajectory Brier do not strictly elicit the success-probability trace.
- Introduced the Trajectory Proper Score (TPS), strictly proper trajectory-level scoring rules built from any strictly proper binary score; proved strict elicitation and extended it to step-budget censored runs by projecting onto the observed prefix.
- Across four benchmarks, evaluator choice changes conclusions: on Tau2-Bench, AUROC barely separates raw and recalibrated confidence (**0.3 SE**) while TPS shows the raw stream is worse than a base-rate guess (**43 SE**); on WebShop, complete-only evaluation drops the **47%** of runs that hit the step budget, and censored-aware scoring shifts the verdict (95% CI excludes zero).

SELFDOUT: Uncertainty from Reasoning Traces (ICML-W [2])

[Paper]

- Asked whether one reasoning trace can signal uncertainty when APIs expose no logits.
- Proposed the Hedge-to-Verify Ratio (doubt expressed vs. doubt checked), with model-specific markers discovered from 90 unlabeled traces, fused with verbalized confidence; needs no training or model internals.
- Across **7 models and 3 benchmarks**, zero-hedge traces are correct **96%** of the time at 25% coverage; the full score beats semantic entropy on AUROC ($p = 0.001$) at **10× lower cost**, and a deferral cascade reaches 89.7% accuracy at 71% coverage.

Evidence Collapse in Multimodal Reasoning (COLM-W [3])

[Paper]

- Found that reasoning VLMs lose grounding as they think: evidence attention falls **53–91%** in **all nine model–dataset settings**, invisible to text-only entropy.
- Showed the risk is task-conditional: confident-but-blind predictions are hazardous on visual-reference tasks but benign on symbolic ones; a targeted vision veto cuts selective risk by up to **1.9 pp at 90% coverage**, helping exactly where predicted.

TECHNICAL SKILLS

Languages & Frameworks: Python, C++; PyTorch, Transformers.

Training & Inference: FSDP2, tensor parallelism (DTensor), QLoRA, vLLM, Ray.

Environments & Tools: Harbor, Prime Env, Docker, Redis.