

SATWIK PANDEY

[mail](#) | [scholar](#) | [website](#) | [linkedin](#) | [github](#)

EDUCATION

Jaypee Institute of Information Technology

Bachelor of Technology in Computer Science; CGPA: 3.3/4.0

Noida, India

Aug 2021 – May 2025

RESEARCH INTERESTS

- **Primary Areas:** NLP · LLMs · Uncertainty Quantification · Reliability · Reasoning · Trustworthy AI · AI Safety/Security

PUBLICATIONS & PREPRINTS

- [1] Suresh Raghu*, **Satwik Pandey***, Shashwat Pandey. "Proper Scoring Rules for Agentic Uncertainty Quantification." *arXiv preprint. ICML 2026 CTB, ICML 2026 FAGEN.*
 - Introduced a family of **strictly proper** trajectory level scoring rules for evaluating **uncertainty in LM agents**, with a censored-trace extension and negative results establishing that standard trajectory level calibration metrics (ECE, Brier) are not strictly proper in the agentic setting
- [2] **Satwik Pandey***, Suresh Raghu*, Shashwat Pandey. "SELFDOUT: Uncertainty Quantification for Reasoning LLMs via the Hedge-to-Verify Ratio." *arXiv preprint. ICML 2026 FAGEN.*
 - Proposed an **O(1) black-box** uncertainty framework that extracts behavioral hedge/verify signals from reasoning traces, significantly outperforming **Semantic Entropy** on discrimination ($p=0.001$) at **10× lower cost**; a zero-hedge gate achieves **96.1% precision** across 7 models and 3 benchmarks.
- [3] Suresh Raghu*, **Satwik Pandey***. "Don't Blink: Evidence Collapse during Multimodal Reasoning." *arXiv preprint. Under Review*
 - Identified a universal **evidence collapse** phenomenon in reasoning VLMs, observing visual attention drops up to **90.8%** during generation and discovering a **task-conditional failure regime** where confident but visually disengaged predictions are hazardous on sustained visual reference tasks but benign on symbolic tasks.
- [4] **Satwik Pandey**, et al. "Repair of Thought: Advancing Automated Program Repair through a Dual-Model Reasoning Framework." *Manuscript under review at Journal of Systems and Software.*
 - Introduced a function-level APR framework achieving a **SOTA 83.1% plausible repair rate** on Defects4J, with an automated verification pipeline combining AST alignment, control-flow symbolic analysis, and semantic checks.

RESEARCH EXPERIENCE

University of California, Santa Cruz (AIEA Lab)

Research Assistant | Advisor: Prof. Leilani H. Gilpin

Remote / Santa Cruz, CA

Jun 2024 – Aug 2024

- Investigated **Trustworthy Logical Reasoning** in LLMs by translating natural language outputs into Prolog-based logic constraints to mathematically verify consistency and detect hallucinations.
- Designed a **Modular RAG** architecture to decouple knowledge retrieval from reasoning, reducing false positive hallucination detection by integrating external ground-truth validation.

Jaypee Institute of Information Technology

Undergraduate Researcher | Advisor: Prof. Sandeep Kumar Singh

Noida, India

Aug 2023 – May 2024

- Conducted initial experiments on **parameter-efficient fine-tuning** (Unsloth/QLoRa) which revealed limitations in complex logic repair, motivating a strategic pivot from training-based optimization to inference-time reasoning framework that became the core contribution of the "Repair of Thought" publication.

PROFESSIONAL EXPERIENCE

VFS Global

AI Research Engineer

New Delhi, India

Dec 2024 – Present

- Investigated dynamic execution paths for **multi-agent frameworks**. Designed a **complexity-based routing mechanism** that differentiates between deterministic tasks (handled by parsers) and ambiguous instances (routed to reasoning-capable VLMs), optimizing the trade-off between inference cost and reasoning accuracy.
- Integrated a two-stage confidence pipeline: **token-level logit probabilities** aggregated via **Shannon entropy** serve as a lightweight first-pass filter for extraction uncertainty, with high-entropy predictions escalated to a **reasoning VLM-as-a-Judge** for **cross-modal extraction verification**.
- Designed a **trace-guided re-extraction** mechanism for cascade rejected fields that fail both entropy screening and judge validation. The judge's **reasoning traces** are converted into corrective context for a secondary extraction model, enabling targeted recovery from extraction errors such as character misreads, boundary confusion, and semantic misattribution.
- Designed a **multi-level verification framework** for biometric photo validation, separating **signal-level quality assessment** from **semantic-level attribute verification** with fine-tuned VLMs. The staged design isolated distinct failure modes and improved the **interpretability** of rejection decisions.