Satwik Pandey

psatwik2711@gmail.com | satwik-pandey.com | linkedin.com/in/satwik-pandey | github.com/satwik2711

Education

Jaypee Institute of Information Technology — Noida, India
Bachelor of Technology in Computer Science; CGPA: 3.3/4.0 — *Aug 2021 – May 2025*

Research Interests

- **Primary Areas:** Natural Language Processing · Large Language Models · Uncertainty Quantification · Trustworthy AI · AI Safety/Security

Publications & Preprints

S. Pandey, et al. "SELFDOUBT: Uncertainty Quantification for Reasoning LLMs via the Hedge-to-Verify Ratio."
arXiv preprint. Submitted to COLM 2026.

- 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**$\times$ **lower cost**; a zero-hedge gate achieves **96.1% precision** across 7 models and 3 benchmarks.

S. Raghu, **S. Pandey**. "Don't Blink: Evidence Collapse during Multimodal Reasoning."
arXiv preprint. Submitted to UAI 2026.

- Identified a universal **evidence collapse** phenomenon in reasoning VLMs, observing visual attention drops up to **90.8%** during generation and discovered a **task-conditional failure regime** where confident but visually disengaged predictions are hazardous on sustained visual reference tasks but benign on symbolic tasks.

S. 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 an **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) — Remote / Santa Cruz, CA
Research Assistant | *Advisor: Prof. Leilani H. Gilpin* — *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.
- Developed an automated evaluation pipeline to benchmark trustworthiness, contributing to an open-source framework for explainable error reasoning in AI systems.

Jaypee Institute of Information Technology — Noida, India
Undergraduate Researcher | *Advisor: Prof. Sandeep Kumar Singh* — *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 — New Delhi, India
AI Research Engineer — *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.