Sanchit Kabra

sanchit23@vt.edu Sanchit Kabra Sanchit-404

Research Focus

Self-Evolving Agents; Reasoning in Large Language Models; Red Teaming of LLM Agents; AI for Scientific Discovery

Education

Virginia Tech **2024 – Present**
PhD, MS (exp. 2026) in Computer Science *Virginia, USA*

Birla Institute of Technology and Science, Pilani **2020 – 2024**
Bachelor of Engineering in Computer Science (Honors) *Pilani, India*

Publications *Google Scholar*

- **Accelerating Materials Design via LLM-Guided Evolutionary Search**. Nikhil Abhyankar*, **Sanchit Kabra***, Saaketh Desai, Chandan K. Reddy. *ICLR 2026.*
- **SURFACEBENCH: Can Self-Evolving LLMs Find the Equations of 3D Scientific Surfaces?**. Sanchit Kabra, Shobhnik Kriplani, Parshin Shojaee, Chandan K. Reddy. *TMLR*, 2026.
- **Decompose, Adapt and Evolve: Towards Efficient Scientific Equation Discovery with Large Language Models**. Pouya Behzadifar*, Parshin Shojaee*, **Sanchit Kabra***, Kazem Meidani, Chandan K. Reddy. *NeurIPS AI for Science Workshop*, 2025; *ICML*, 2026.

Preprints

- **Reasoning Towards Fairness: Mitigating Bias in Language Models through Reasoning-Guided Fine-Tuning**. **Sanchit Kabra**, Akshita Jha, Chandan K. Reddy. *Preprint*, 2025.
- **Biased or Flawed? Mitigating Stereotypes in Generative Language Models by Addressing Task-Specific Flaws**. Akshita Jha, **Sanchit Kabra**, Chandan K. Reddy. *Preprint*, 2024.
- **RUST-BENCH: Benchmarking LLM Reasoning on Unstructured Text within Structured Tables**. Nikhil Abhyankar, Purvi Chaurasia, **Sanchit Kabra**, Ananya Srivastava, Vivek Gupta, Chandan K. Reddy. *Preprint*, 2025.
- **Aligned at the Start: Concept Formation in LLM Embeddings**. Mehrdad Khatir, **Sanchit Kabra**, Chandan K. Reddy. *Preprint*, 2024.

Research Experience

Amazon Science – *Applied Science Intern*, Stores, Economics and Science **May 2026 – August 2026**
Hypothesis Generation for Demand Substitutability *Palo Alto, CA*

- Developed an automated hypothesis generation system leveraging customer purchase trajectories to surface latent hypotheses for answering abstract business questions and driving weblab experiments at scale.
- Built an end-to-end auto-research framework to improve demand substitution model along axes of model architecture and hyperparameter optimization.

Virginia Tech – *Research Assistant* **May 2022 – Present**
Advisor: Chandan Reddy *Arlington, VA*

Red Teaming & Security Evaluation

- **LLMs as Evolutionary Attackers:** Developed a novel LLM-driven evolutionary framework for discovering single-turn jailbreak prompts under strict black-box query budgets.
  - Discovered **3× more** unique successful jailbreaks than template-based attacks under the same budget.
  - Achieved **100% ASR** and **82–86%** judge-validated ASR on LLaMA-2-7B-Chat and GPT-4o-mini, surfacing structurally distinct attack strategies.
- **Multi-Vulnerability Code Security Benchmarking:** Developed a novel file-level benchmark for multi-vulnerability detection across C, C++, Python, and JavaScript.
  - Generated long-context programs with multiple concurrent CWEs per file; identified severe performance degradation with recall dropping by **30–50%** and exact-file accuracy decreasing by up to **40%** in dense vulnerability settings.

AI for Science

- **Materials Discovery (LLEMA):** Developed an LLM-guided evolutionary search system for multi-objective materials discovery evaluated across 10 industrial-scale tasks.

- Achieved up to **2× higher hit-rates** than generative and LLM-only baselines, consistently discovering chemically valid and thermodynamically stable candidates.
- Produced stronger Pareto fronts with higher hypervolume and earlier convergence across all tasks.

- **Scientific Equation Discovery (DecAEvolve):** Developed DecAEvolve, a novel adaptive reinforcement learning based evolutionary framework for LLM-based symbolic regression that adapts the hypothesis generator at test time.
  - Reduced discovery error by **1.5–3×** (normalized MSE) compared to the state-of-the-art LLM-SR baseline under identical 3K-query budgets.
  - Achieved **1.5–2× faster convergence**, reaching comparable or lower error in **40–60% fewer iterations** across 0.5B and 1.5B models.

Responsible AI & Alignment

- **Bias Analysis and Mitigation in Generative Language Models:** Designed a framework to disentangle identity-conditioned bias from task-specific comprehension and reasoning failures in ambiguous language understanding.
  - Found that **>80%** of stereotypical errors arise from reasoning failures rather than inherent bias; reduced stereotypical outputs by **60%+** through instruction tuning without fairness-specific data.
- **Reasoning-Guided Bias Mitigation:** Developed a reasoning-guided fine-tuning approach to distill structured reasoning traces into non-reasoning models, improving fairness accuracy from **∼40–50%** to **∼85–90%** across multiple fairness dimensions.

Projects

Amazon Trusted AI Challenge – Code-Based LLMs **October 2024 – June 2025**
Advisor: Dr. Ruoxi Jia *Amazon / Virginia Tech*

- Developed an 8B-parameter code-focused LLM from scratch, including pretraining and finetuning, as part of one of the 5 blue teams at the Amazon Trusted AI Challenge, ensuring robustness in code generation tasks.
- Engineered a Chain-of-Thought knowledge distillation framework to transfer reasoning capabilities from GPT-4-O1 and DeepSeek R1 to an 8B parameter base model, improving performance on complex programming tasks.
- Designed adversarial and supervised finetuning pipelines to reduce the model's susceptibility to generating malicious, insecure, or vulnerable code while maintaining functional correctness.
- Conducted large-scale evaluations on security benchmarks using static analysis tools and dynamic LLM-driven systems for context-specific vulnerability assessment.

Awards

- **Amazon Trusted AI Challenge**, 2024: A global university competition focused on responsible AI and LLM coding security. Part of **HokieTokie**, one of 10 teams selected globally. Received **USD 900,000** in AWS compute credits over the course of the competition. Advisor: Ruoxi Jia.
- **Higher Secondary School Exams State Topper**, 2020: Ranked second in the Maharashtra State Board Examinations among 1.5 million students. Scored a perfect 100/100 in Math, Chemistry, Physics, and Biology.
- **INSPIRE Scholarship**, 2020: Offered by the Department of Science & Technology (DST), Government of India for finishing in the top 1% in Class XII Board Examinations.

Skills

- **Programming Languages:** Python, Java, R, Matlab, C/C++, SQL, bash
- **Frameworks and Tools:** PyTorch, TensorFlow, Numpy, Pandas, RAG, Huggingface, SciKit, Matplotlib, pyspark
- **Technologies:** AWS, Linux

Relevant Coursework

- Data Structures and Algorithms
- Design and Analysis of Algorithms
- Operating Systems
- Database Systems
- Compiler Construction
- Computer Architecture
- Deep Learning
- Natural Language Processing
- Multimodal Vision
- Decision Making Under Uncertainty
- Computational Learning Theory