# Gaurav Srivastava

☎ +1 (540) 934-8111 ✉ gks@vt.edu LinkedIn GitHub Google Scholar Kaggle (3X Expert) Website

## Education

**Virginia Tech University** Blacksburg, Virginia
Master of Science in Computer Science (Fully Funded - **62,705$/year** Scholarship), **GPA:** 4.0/4.0 Aug 2024 - May 2026 (Expected)

- **Advisor:** Dr. Xuan Wang; **Thesis Committee:** Dr. Tu Vu, Dr. Naren Ramakrishnan, Dr. Chris Thomas
- Graduate Teaching Assistant for CS5834 (Fall 2025), CS5814 (Spring 2025), CS1064 (Fall 2024)

**Manipal University Jaipur** Jaipur, India
Bachelor of Technology in Computer Science and Engineering, **GPA:** 9.10/10.0 Jul 2019 – Jul 2023

## Experience

**Dell Technologies - Office of the CTO (OCTO)** May 2025 - Aug 2025
*AI Research Intern* Austin, Texas

- Architected autonomous resource allocation system using **11 specialized AI agents** with **57 tools**, improving GPU utilization from **8→40%**, achieving ~**25% cost reduction** and **35-40% better decision quality**.
- Deployed production system on real PowerEdge server fleets, processing **1000+ concurrent workloads** with **89% cost efficiency**, **91% success rate**, and **26.5% improvement** in decision quality over Kubernetes/SLURM schedulers.
- Built algorithm lifecycle management system with **4 AI agents** enabling autonomous selection, extraction, validation, and **zero-downtime replacement** of production algorithms from **academic papers** via Semantic Scholar/arXiv APIs.

*Submitted **4 patents**; Published internal paper *OCTO-11136: Towards an Agentic Approach to Autonomous Resource Allocation*

**Dell Technologies** Aug 2023 - Jul 2024
*Machine Learning Engineer* Hyderabad, India

- Developed **DDS-GPT**, a RAG-based tool using flan-t5-large and instructor-xl embeddings that utilizes Dell Design System docs to generate code snippets for UI components, saving UI developer's manual efforts by ~**60%**.
- Automated metrics monitoring dashboard for **59 product health metrics** (e.g., CI/CD maturity), saving ~**4 days** per sprint for every product manager in eCommerce Org by cutting report generation from **4 days** → <**15 minutes**.
- Fine-tuned BERT-based error classification models on Splunk error logs (**97.39% F1**); then optimized to ML ensemble (DT, RF, XGBoost) with similar accuracy (drop=<2%), cutting inference time from **3 mins→16 sec for 1M records** and removing GPU dependency; reducing manual efforts by **80%** and boosting job success rates by **24%** under EBI Org.
- Led the adoption of MLOps within Dell's ecommerce Org, automating ML model monitoring and retraining processes.

## Selected Publications

- **G. Srivastava,** Shuxiang Cao, and Xuan Wang. "ThinkSLM: Towards Reasoning in Small Language Models." in *Proc. 2025 Conf. of Empirical Methods in Natural Language Processing* **(EMNLP'25 Main)**. **arxiv** | **leaderboard**
- **G. Srivastava,** Zhenyu Bi, Meng Lu, and Xuan Wang. "DEBATE, TRAIN, EVOLVE: Self-Evolution of Language Model Reasoning." in *Proc. 2025 Conf. of Empirical Methods in Natural Language Processing* **(EMNLP'25 Main)**. **arxiv**
- **G. Srivastava,** Aafiya Hussain, Zhenyu Bi, et al. (+5 authors). "BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models." **(Under Review in ICLR 26)** arXiv:2509.24210. **arxiv** | **leaderboard**
- **G. Srivastava,** Aafiya Hussain, Sriram Srinivasan, and Xuan Wang. "Do LLMs Overthink Basic Math Reasoning? Benchmarking the Accuracy-Efficiency Tradeoff in Language Models." **(Under Review in ACL 26)** arXiv:2507.04023. **arxiv**

***Complete list of publications** - **Google Scholar**, **total publications:** 24, **citations:** 208, **h-index:** 8

## Selected Projects

**LLMThinkBench (4.11K+ PyPI downloads) | Python | vLLM | Transformers** **GitHub** | **PyPI** | **Leaderboard** Apr 2025

- Benchmark framework evaluating LLM reasoning across **14+ tasks** with **pass@k** evaluation, multi-GPU inference via vLLM, and novel **Overthinking Score** metric balancing accuracy with token efficiency using F1-harmonic mean.
- Achieved **500+ samples/task** reproducibility with modular architecture supporting custom task extensions, standardized prompt templates, and comprehensive metrics including instruction-following rates and token analysis.

**DataSense - Multi-Agent Data Visualization | Python | Streamlit | vLLM | Plotly** **GitHub** Apr 2025

- Built a visualization system with **3+ agent ensemble** using consensus voting to recommend top **3 chart types** from **9+ options**, auto-generating Plotly visualizations and data narratives with **75%** faster analysis vs manual exploration.

## Technical Skills

**Programming languages:** Python, C, C++ **Lib/Frameworks:** Pytorch, TensorFlow, vLLM, sklearn, Langchain
**Technologies:** Flask, Elasticsearch, MySQL **Tools:** Databricks, AWS, Sagemaker, FastAPI, git, gitlab, Docker

## Honors & Awards

- 3 Inspire Recognition Awards for Innovation and positioning Dell PowerEdge as "AI-native" infrastructure, Dell 2025
- President's Gold Medal Award for Excellence in Research, Manipal University 2023
- Five-time recipient of the Dean and Student Excellence Awards for publishing research, Manipal University 2022-2023
- 3 times All India Grand Finalist - Wipro GE Healthcare, NEC and Mitsubishi, and T-Systems Hakcathon 2022
- Winner, NPSiHacks **(Project - AI Verifica)** 2021