

# GAURAV SRIVASTAVA

+1 (540) 934-8111 [✉ gks@vt.edu](mailto:gks@vt.edu) [in LinkedIn](#) [🌐 GitHub](#) [🎓 Google Scholar](#) [📄 Kaggle \(3X Expert\)](#) [🌐 Website](#)

## EDUCATION

### Virginia Tech University

Master of Science in Computer Science (Fully Funded), **GPA: 4.0/4.0**

Blacksburg, Virginia

Aug 2024 - May 2026

- **Thesis:** Enabling Small Language Models as Efficient and Capable Agents
- **Advisor:** [Dr. Xuan Wang](#); **Thesis Committee:** [Dr. Xuan Wang](#), [Dr. Naren Ramakrishnan](#), [Dr. Chris Thomas](#), [Dr. Tu Vu](#)
- Graduate Teaching Assistant for CS5624 (Spring 26), CS5834 (Fall 25), CS5814 (Spring 25), CS1064 (Fall 24)

### Manipal University Jaipur

Bachelor of Technology in Computer Science and Engineering, **GPA: 9.10/10.0**

Jaipur, India

Jul 2019 – Jul 2023

## EXPERIENCE

### Dell Technologies - Office of the CTO (OCTO)

*AI Research Intern*

May 2025 - Aug 2025

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

*Machine Learning Engineer*

Aug 2023 - Jul 2024

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 the entire Product Experience Platform (PXP) CDO metrics dashboard for **59 product health metrics**, saving **~4 days** per sprint for every product manager in Dell's eCommerce Org.
- Worked with cross-functional teams to improve a Gedis class predictor, resulting in a **12% increase in SKU accuracy**.
- Engineered a customer support bot using Dell-hosted fine-tuned **Llama2-70B** on Jira Data to dynamically route incident tickets to appropriate SMEs, reducing **40% of SLA times**.
- Led the adoption of MLOps within Dell's eCommerce Org, enhancing ML models monitoring and automating retraining.
- Built a dynamic forecasting system for the Product Data team to alert 59 clients, saving infra cost by **~40%**.

### Dell Technologies

*SDE Intern (Machine Learning)*

Feb 2023 - May 2023

Hyderabad, India

- Developed ML capabilities to analyze and extract precise error patterns within daily failed Job Logs for the Enterprise Business Intelligence (EBI) team, contributing to enhanced data-driven insights and decision-making processes.
- Fine-tuned **RoBERTa** using Splunk Tables' generated Error Logs data, achieving an F1-score of **98.03%** with inference time of 22 min 36s for 10 million records.
- Devised a combined model pipeline integrating Decision Trees, Naive Bayes, and XGBoost, strategically dividing data based on best-performing model for each cluster, reducing inference time to **3 min 41s** for 10M records.
- Designed end-to-end ML ingestion pipeline to monitor and retrain the model whenever failure jobs deviated from existing cluster scope, ensuring continuous model relevance and accuracy.
- Deployed and operationalized the model on Pivotal Cloud Foundry (PCF), optimizing processing speed to **18ms per record** while handling batches of 1000 records in 3.2s.

### Swiggy (Bundl Technologies)

*Applied Research Intern (Computer Vision)*

Sep 2022 - Jan 2023

Bengaluru, India

- Improved Swiggy's image similarity model using supervised learning on **Barlow Twins** embeddings, achieving **4.51%** accuracy boost on validation and **12-13%** on test sets.
- Engineered a streamlined CV pipeline using Streamlit for Swiggy Dine-out, generating advertisement videos from static restaurant and food images.
- Automated the manual embeddings generation process for Swiggy Supr Daily reducing manual efforts.
- Collaborated on the Swiggy x IIT Jodhpur project, exploring object detection models for facial verification.

### Dell Technologies

*SDE Intern (Data Science & Automation)*

Jun 2022 - Sep 2022

Hyderabad, India

- Worked with the Enterprise Business Intelligence (EBI) team to automate and enhance data reliability processes, streamlining data workflows and ensuring data integrity.
- Developed a feature for D'Owl tool to remove feature-level duplicates from records, enhancing data quality and reducing redundancy.
- Optimized the D'Owl codebase, reducing processing time from initial **7.68 milliseconds** to **997 microseconds** for 1 record.
- Explored NLP capabilities to extract valuable information from emojis rather than simply removing them from records.

- **Gaurav Srivastava**, Aafiya Hussain, Chi Wang, Yingyan (Celine) Lin, and Xuan Wang. “effGen: Enabling Small Language Models as Capable Autonomous Agents.” in *Proc. Forty-third International Conference on Machine Learning (ICML’26)*. [arxiv](#) [effgen.org](#) | [docs](#)
- **Gaurav Srivastava**, Aafiya Hussain, Zhenyu Bi, Swastik Roy, Priya Pitre, Meng Lu, Morteza Ziyadi, and Xuan Wang. “BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models.” in *Proc. The 14th International Conference on Learning Representations (ICLR’26)*. [openreview](#) | [arxiv](#) | [leaderboard](#)
- **Gaurav 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)*. [acl](#) | [leaderboard](#)
- **Gaurav 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)*. [acl](#) | [website](#)
- **Gaurav Srivastava**, Aafiya Hussain, Sriram Srinivasan, and Xuan Wang. “Do LLMs Overthink Basic Math Reasoning? Benchmarking the Accuracy-Efficiency Tradeoff in Language Models.” in *Proc. 64th Annual Meeting of the Association for Computational Linguistics (ACL’26 Findings)*. [arxiv](#) | [leaderboard](#)
- Zhenyu Bi\*, **Gaurav Srivastava\***, Yang Li, Swastik Roy, Meng Lu, Morteza Ziyadi, and Xuan Wang. “JudgeBoard: Benchmarking and Enhancing Small Language Models for Reasoning Evaluation.” in *Proc. The 40th Annual AAAI Conference on Artificial Intelligence (AAAI’26 Oral)*. [arxiv](#)
- Aafiya Hussain, **Gaurav Srivastava**, Alvi Ishmam, Zaber Hakim, and Chris Thomas. “SoundBreak: A Systematic Study of Audio-Only Adversarial Attacks on Trimodal Models.” in *Proc. 64th Annual Meeting of the Association for Computational Linguistics (ACL’26 Main)*.
- Priya Pitre, **Gaurav Srivastava**, Lu Zhang, Le Wang, Naren Ramakrishnan, and Xuan Wang. “A Diagnostic Study of Multi-Agent LLMs for Real-World Debates.” in *Proc. Forty-third International Conference on Machine Learning (ICML’26)*.
- Meng Lu, Ran Xu, Yi Fang, Wenxuan Zhang, Yue Yu, **Gaurav Srivastava**, Yuchen Zhuang, Mohamed Elhoseiny, Charles Fleming, Carl Yang, Zhengzhong Tu, Yang Xie, Guanghua Xiao, Hanrui Wang, Di Jin, Wenqi Shi, and Xuan Wang. “Scaling Agentic Reinforcement Learning for Tool-Integrated Reasoning in VLMs.” in *Proc. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR’26)*. [arxiv](#)
- Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, **Gaurav Srivastava**, Xuan Wang, and Naren Ramakrishnan. “Hindsight: Structured Agent Memory that Retains, Recalls, and Reflects.” in *Proc. 64th Annual Meeting of the Association for Computational Linguistics (ACL’26 System Demo)*. [openreview](#) | [pypi](#) | [github](#) | [website](#) | [vectorize.io](#)
- Priya Pitre, **Gaurav Srivastava**, Lu Zhang, Le Wang, Naren Ramakrishnan, and Xuan Wang. “Beyond Consensus: Evaluating Multi-Agent LLM Debates through a Deliberative Democracy Framework.” in *Proc. Generation, Evaluation & Metrics Workshop (GEM @ ACL’26)*. [openreview](#)
- Chris Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, **Gaurav Srivastava**, Xuan Wang, and Naren Ramakrishnan. “Hindsight is 20/20: Building Agent Memory that Retains, Recalls, and Reflects.” [arxiv](#)
- Andrew Neeser, **Gaurav Srivastava**, Kaylen Latimer, Aadyant Khatri, Xuan Wang, Christopher Latimer, and Naren Ramakrishnan. “QuOTE: Question-Oriented Text Embeddings.” (U.R. in KDD’26).
- Meng Lu, Yuchen Zhuang, Wenqi Shi, **Gaurav Srivastava**, Charles Fleming, and Xuan Wang. “MAF-IE: Multi-Agent Finetuning for Zero-shot Information Extraction.”
- Priya Pitre, **Gaurav Srivastava**, Lu Zhang, Le Wang, Naren Ramakrishnan, and Xuan Wang. “SIMAGENT: Towards Multi-Agent LLM for Real-World Stakeholder Debate Simulations without Ground Truth.”
- **Gaurav Srivastava\***, Meng Lu\*, and Xuan Wang. “Towards Small (Vision-)Language Models as the Future of Real-World Agents.”

## OLDER PUBLICATIONS (BEFORE 2025)

- **Gaurav Srivastava** and Nitesh Pradhan. “Handling imbalanced class in melanoma: Kemeny-Young rule based optimal rank aggregation and Self-Adaptive Differential Evolution Optimization.” *Engineering Applications of Artificial Intelligence*, 2023. (IF: 8.0) [link](#)
- **Gaurav Srivastava**, Aninditaa Chauhan, and Nitesh Pradhan. “CJT-DEO: Condorcet’s jury theorem and differential evolution optimization based ensemble of deep neural networks for pulmonary and colorectal cancer classification.” *Applied Soft Computing*, 2023. (IF: 8.3) [link](#)
- **Gaurav Srivastava**, Nitesh Pradhan, and Yashwin Saini. “Ensemble of deep neural networks based on Condorcet’s jury theorem for screening COVID-19 and pneumonia from radiograph images.” *Computers in Biology and Medicine*, 2022. (IF: 7.7) [link](#)
- **Gaurav Srivastava**, Aninditaa Chauhan, Nitigya Kargeti, Nitesh Pradhan, and Vijaypal Singh Dhaka. “ApneaNet: A hybrid 1DCNN-LSTM architecture for detection of Obstructive Sleep Apnea using digitized ECG signals.” *Biomedical Signal Processing and Control*, 2023. (IF: 5.1) [link](#)
- **Gaurav Srivastava**, Aninditaa Chauhan, Mahesh Jangid, and Sandeep Chaurasia. “Covixnet: A novel and efficient deep learning model for detection of COVID-19 using chest X-ray images.” *Biomedical Signal Processing and Control*, 2022. (IF: 5.1) [link](#)
- Amitesh Kumar Dwivedi, **Gaurav Srivastava**, Sakshi Tripathi, and Nitesh Pradhan. “eFuseNet: A deep ensemble fusion network for efficient detection of Arrhythmia and Myocardial Infarction using ECG signals.” *Multimedia Tools and Applications*, 2024. (IF: 3.0) [link](#)

- Nitesh Pradhan, **Gaurav Srivastava**, and Geetika Kaushik. “Vit-Ensemble: Probabilistic voting based ensemble of Vision Transformers for tuberculosis detection using radiographs.” *Computational Biology and Chemistry*, 2024. [link](#) [↗](#)
- **Gaurav Srivastava** and Mahesh Jangid. “Multi-view sparse laplacian eigenmaps for nonlinear spectral feature selection.” *2023 International Conference on System Science and Engineering (ICSSE)*, IEEE, 2023. [link](#) [↗](#)
- Amitesh Kumar Dwivedi, **Gaurav Srivastava**, and Nitesh Pradhan. “NFF: A novel nested feature fusion method for efficient and early detection of colorectal carcinoma.” *Proceedings of Fourth International Conference on Computer and Communication Technologies*, Springer, 2023. [link](#) [↗](#)
- Ayush Singh, **Gaurav Srivastava**, and Nitesh Pradhan. “Pneumothorax segmentation using feature pyramid network and MobileNet encoder through radiography images.” *International Conference on Soft Computing and Signal Processing*, Springer, 2023. [link](#) [↗](#)
- Rugved Sanjay Chavan, **Gaurav Srivastava**, and Nitesh Pradhan. “Advance plant health monitoring and forecasting system using edge-fog-cloud computing and LSTM networks.” *Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications*, Springer, 2023. [link](#) [↗](#)
- Nitesh Pradhan, Saransh Gupta, and **Gaurav Srivastava**. “Image colorization: A convolutional network approach.” *Proceedings of International Conference on Data Science and Applications*, Springer, 2023. [link](#) [↗](#)
- **Gaurav Srivastava**, Aninditaa Chauhan, Mahesh Jangid, and Ashish Jain. “An analysis of deep learning models to diagnose COVID-19 using radiography images.” *2022 International Conference for Advancement in Technology (ICONAT)*, IEEE, 2022. [link](#) [↗](#)
- **Gaurav Srivastava**, Devika Sapra, Akruhi Sinha, Mahin Anup, and Deepak Sinwar. “Artificial intelligence and IoT-assisted sustainable manufacturing for Industry 4.0.” *Computational Intelligence based Optimization of Manufacturing Process for Sustainable Materials*, CRC Press, 2024. [link](#) [↗](#)
- Akruhi Sinha, **Gaurav Srivastava**, Devika Sapra, and Chhavi Deshlahra. “Fog computing for agriculture applications and its issues.” *Bio-Inspired Optimization in Fog and Edge Computing Environments*, Auerbach Publications, 2023. [link](#) [↗](#)
- Akruhi Sinha, Deepak Sapra, **Gaurav Srivastava**, Mahin Anup, and Deepak Sinwar. “AI-assisted big data analytics for smart healthcare systems.” *Intelligent Internet of Things for Smart Healthcare Systems*, CRC Press, 2023. [link](#) [↗](#)

[Google Scholar](#) [↗](#) | **Total:** 32 publications (17 during Virginia Tech), **Citations:** 309, **h-index:** 10

## SELECTED PROJECTS

**effGen (8K+ PyPI downloads, 150+ GitHub stars) | Python | vLLM**    [GitHub](#) [↗](#) | [PyPI](#) [↗](#) | [effgen.org](#) [↗](#) | [Docs](#) [↗](#)    Oct 2025

- Developed open-source agentic framework optimized for **Small Language Models**, achieving **100%** accuracy on hard reasoning tasks vs **13.3%** for LangChain/AutoGen (**~8x improvement**), with **3.7x faster** inference via native vLLM.
- Engineered **15+ built-in tools**, MCP/A2A/ACP protocol support, automatic task decomposition, and multi-GPU tensor parallelism, enabling SLMs to match LLM-level performance on complex multi-step reasoning benchmarks.

**BeyondBench (1.3K+ PyPI downloads) | Python | vLLM | Transformers**    [GitHub](#) [↗](#) | [PyPI](#) [↗](#) | [Leaderboard](#) [↗](#)    Sep 2025

- Developed dynamic evaluation framework with **29 programmatically-generated tasks** across **3 difficulty levels** (Easy: 14, Medium: 5, Hard: 10), eliminating benchmark contamination by generating **near infinite unique test instances**.
- Evaluated **101 LLMs** including GPT-4, Gemini, Llama, and Qwen families; revealed **40-60%** accuracy drop on hard constraint-satisfaction tasks (Sudoku, N-Queens, SAT) compared to standard benchmarks.

**LLMThinkBench (5.4K+ 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

## TEACHING

- **GTA**, Natural Language Processing (CS5624), Virginia Tech - Dr. Xuan Wang    Spring 2026
- **GTA**, Introduction to Urban Computing (CS5834), Virginia Tech - Dr. Naren Ramakrishnan    Fall 2025
- **GTA**, Introduction to Deep Learning (CS5814), Virginia Tech - Dr. Xuan Wang    Spring 2025
- **GTA**, Introduction to Programming in Python (CS1064), Virginia Tech - Dr. John Lewis    Fall 2024
- **AI/ML Mentor**, [Edu-versity](#) - Designed AI course with **4.9/5.0** rating, enrolled by **6,500+** students    May-Aug 2022

## SERVICE

- **Program Committee (Reviewer):** AAAI 2026, ICDM 2025, KDD 2024
- **Journal Reviewer:** IEEE Transactions on Neural Networks and Learning Systems, IEEE Access (30+ articles), Springer Applied Intelligence, Springer Multimedia Systems, Springer JHIR, PLOS ONE

## HONORS & AWARDS

---

- Outstanding MS Research Award, Virginia Tech CS Department 2026
- [Torgersen Graduate Research Excellence Award](#) Finalist (Top 10), Virginia Tech 2026
- \$50K Commonwealth Cyber Initiative (CCI) Grant for [effGen](#), a B2B SaaS agent-building framework startup 2026
- Accepted to NSF I-Corps™ Program, a 7-week entrepreneurial training program for commercialization of [effGen](#) 2026
- 3 Inspire Recognition Awards for positioning Dell PowerEdge as "AI-native" infrastructure, Dell Technologies 2025
- President's Gold Medal Award for Excellence in Research, Manipal University Jaipur 2023
- Runner-up, Dell IT Development Program (ITDP) FY'23 Hackathon, Dell Technologies 2023
- Ranked 13/473 globally in Bitgrit Generative AI Competition; 117/26,008 in Amazon ML Challenge 2023 2023
- Three-time recipient of Student Excellence Award for publishing research, MUJ; Best Research Project, CS Dept 2022-23
- 3 times All India Grand Finalist - Wipro GE Healthcare, NEC and Mitsubishi, T-Systems Hackathon 2022
- 3rd Position Hack2Hire Hackathon, Dell; Winner NPSiHacks ([AI Verifica](#)) [↗](#); Kaggle 3X Expert 2020-21