SHUBHAM AGARWAL

+91 9083271307 | shagarw@adobe.com | Github | Website | Google Scholar

RESEARCH INTERESTS

My research interests lie in developing efficient designs for enhancing the efficiency of ML systems which encompasses various aspects including scalability, reliability, and optimizations. My research experiences include systems for ML, cloud system reliability, ML for system reliability, and data-driven systems.

EDUCATION

Bachelor of Engineering (B.E.) in Computer Science and Engineering — July 2018 - April 2022
Birla Institute of Technology and Science (BITS) Pilani, Pilani Campus, India — GPA: 9.95/10.00

- Attained the highest GPA across the Institute (Institute Rank - 1)
- Served as student representative in the Student Faculty Council (SFC)
- Awarded 100% merit scholarship for academic excellence by the Institute (awarded to top-1% students)

PUBLICATIONS

[1] (NSDI '24) **Shubham Agarwal**, Subrata Mitra, Sarthak Chakraborty, Srikrishna Karanam, Koyel Mukherjee, Shiv Saini. **Approximate Caching for Efficiently Serving Text-to-Image Diffusion Models.** In *21st USENIX Symposium on Networked Systems Design and Implementation*, 2024. [LINK]

[2] (PAKDD '24) Ghazi Shazan Ahmad, **Shubham Agarwal**, Subrata Mitra, Ryan A. Rossi, Manav Doshi, Syam Manoj Kumar Paila. **ScaleViz: Scaling Visualization Recommendation Models on Large Data.** In *The 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2024.* [LINK] (*Oral*)

[3] (ESEC/FSE '23) **Shubham Agarwal**, Sarthak Chakraborty, Shaddy Garg, Sumit Bisht, Chahat Jain, Ashritha Gonuguntla, Shiv Saini. **Outage-Watch: Early Prediction of Outages using Extreme Event Regularizer.** In *The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 682-694, 2023. [LINK] (Accpt. Rate: 25.6%)

[4] (ASE '23) Sarthak Chakraborty, **Shubham Agarwal**, Shaddy Garg, Abhimanyu Sethia, Udit Narayan Pandey, Videh Aggarwal, Shiv Saini. **ESRO: Experience Assisted Service Reliability against Outages.** In *The 38th IEEE/ACM International Conference on Automated Software Engineering*, 2023. [LINK] (Accpt. Rate: 21.3%)

[5] (WWW '23) Sarthak Chakraborty, Shaddy Garg*, **Shubham Agarwal***, Ayush Chauhan, Shiv Saini. **CausIL: Causal Graph for Instance Level Microservice Data.** In *Proceedings of The Web Conference 2023*, pp. 2905-2915, 2023. [LINK] (Accpt. Rate: 19.2%)

[6] (SIGMOD '23) **Shubham Agarwal**, Gromit Yeuk-Yin Chan, Shaddy Garg, Tong Yu, Subrata Mitra. **Fast Natural Language Based Data Exploration with Samples.** In *Companion of the 2023 International Conference on Management of Data*, pp. 155–158, 2023. [LINK]

INDUSTRIAL RESEARCH

- **Research Associate 2 - Adobe Inc. (BigData Intelligence Lab)** — Jul 2022 - Present
  Group: Systems and Insights Group — *Bangalore, India*
  - Published **6 papers**, filed **2 patents,** and successfully integrated research technologies into **3 products** within 1.5 years
  - Mentored 10 undergraduate interns and 1 PhD intern during summer internships at Adobe over a span of 1.5 years
  - ***ML for System Reliability***
    - **Outage Prediction in Production System (In Product)** [Paper Link]: Implemented the inference and training pipeline of an outage forecasting model by inventing a novel distribution learning approach, exhibiting AUC of 0.8. For the product, we also integrated a Shapley value-based explainability system to pinpoint faulty system alerts.
    - **Root Cause and Remediation Consolidation System** [Paper Link]: Built a diagnostic service for cloud service failures, merging structured alert data and unstructured incident reports data to recommend root causes and remediation efficiently, resulting in a 27% improvement in real-world cloud system diagnosis.
    - **Runtime Prediction of Incoming Jobs (In Product)**: Designed a pipeline for predicting the latency of incoming Spark jobs in a multi-tenant system, constructing features based on system state and incoming job features. It achieves a MAPE value of ~ 0.4 and is internally utilized by developers to assess the need for preventive measures.

- ***System for ML***
  - **Approximate Caching for Diffusion Models (In Product)** [Paper Link]: Built an end-to-end diffusion model serving system and innovated a novel caching technique to reduce text-to-image generation costs by 19% and latency by 19.8% through intelligent reuse of intermediate states without compromising quality.
  - **High-Throughput Text-to-Image Inference Serving** [Paper Link]: Designed a diffusion model inferencing system on a fixed cluster, using *accuracy-scaling* to serve high-quality images in high-load, fixed-budget scenarios. Achieves 10x lower latency SLO violations, 10% higher average quality, and 40% higher throughput.
  - **Scaling VizRec Models on Large Data** [Paper Link]: Devised a novel reinforcement-learning-based scalable framework optimizing input statistics selection for automated visualization recommendation (VizRec) models. Our approach achieves up to 10x speedup in *time-to-visualize* on four large real-world datasets.

RESEARCH INTERNSHIPS

- **Research Intern - American Express (Document Analytics & Intelligence Lab)** Jan 2022 - Jun 2022
  Topic: Document component segmentation and structure analysis for information extraction *Dr. Himanshu S Bhatt*
  - Designed an end-to-end document processing pipeline (using YOLO, LayoutLM) for visually rich documents.
  - Developed a scalable framework for swift adaptation to new document types through automated few-shot learning.
  - Implemented cloud-based inferencing using PyTorch models, and on-device inference using lightweight TFLite models.
  - Received a pre-placement offer for a full-time research role within two months of the start of the internship.
- **Research Intern - Adobe Inc. (BigData Intelligence Lab)** May 2021 - Aug 2021
  Topic: Outage Prediction for Enhanced Cloud System Reliability *Dr. Shiv K Saini*
  - Explored using cloud observability data to predict outages, considering the rarity of these events in real-world data.
  - Developed *OutageWatch*, a novel framework that models outages as extreme events. It predicts these events by forecasting the distribution of QoS metrics and uses a tail-risk regularizer for precise modeling of the distribution tails.
  - Achieved an average AUC of 0.98 and reduced mean time to detect outages by up to 88% for real-world cloud systems.
  - Selected as one of 8 interns among a total of 82, to receive a pre-placement offer for a full-time role in the research team.

PATENTS

[1] *Shubham Agarwal*, Subrata Mitra, Sarthak Chakraborty, Srikrishna Karanam, Koyel Mukherjee, Shiv Kumar Saini. **Intelligent Use of Caching and Retrieval of Intermediate Noise for Resource Efficient Diffusion Models.** [Filing]

[2] Shaddy Garg, *Shubham Agarwal*, Sumit Bisht, Nikhil Sheoran, Chahat Jain, Ashritha Gonuguntla, Shiv Kumar Saini. **A System and Method for Outage Forecasting.** [Filed] (US Patent App. 17/656,263)

TERM PROJECTS

- [Link] Designed an enhanced collaborative filtering recommender system by fine-tuning item weights and similarity scores.
- [Link] Developed an AI tutoring system for teaching algebra, generating questions based on a reward function.
- [Link] Designed grammar rules, implemented a parser and compiler in C for language design and type expression computation.
- [Link] Developed a REACT app for seamless online education, integrating content sharing and video calling features.
- [Link] Developed a lightweight progressive web app client using JavaScript to work with the metastudio.org server.

SKILLS

• **Languages**	Python, C, C++, Java, SQL, Verilog, MIPS
• **Packages and Frameworks**	PyTorch, Keras, TensorFlow, scikit-learn, Kubernetes, Docker, Kafka, Git

RELEVANT COURSES

- **Undergraduate Courses:** Operating Systems, Database Systems, Computer Networks, Compilers, Computer Architecture, Artificial Intelligence, Information Retrieval, Data Structures and Algorithms, Probability and Statistics, Linear Algebra