

Saurabh Gupta

+91 8299819408 | saurabhgupta3012@gmail.com

Visionary AI Leader — Technical Architect — Published Researcher

With 10+ years of experience, I bridge the gap between cutting - edge AI research and enterprise - scale production. From Startups to Fortune 500, I have delivered measurable ROI, including scaling search coverage by 75x, developing AI tools with 1300 DAUs and bringing in \$7.5M business building e2e products.

Experience

Maersk Global

Jan 2026 – Present

AI Platform Engineer

Remote, India

- Creating next generation enterprise platform named **MIDAS** (Maersk Integrated Data and AI Solutions), enabling 10k teams with 100k+ members to use LLMs to build applications in a secure way. Services include an Agent gateway, catalogues, runtime, identity, evaluations and observability etc.
- Developed a control plane to manage authentication and authorization, AI gateway to monitor and govern LLM usage across teams, on prem GPU setup, and SMUS blueprints to handle end to end heavy lifting of modeling process.
- Tech: **GenAI, Platform Engineering, LLM Gateways, Enterprise Security, Governance, Observability, Cloud Platforms, Production Monitoring, AWS Enterprise Agent Framework, Gemini Enterprise Agent Framework**

Cloudsufi Inc. - Google Extended Workforce

Jan 2025 – Dec 2025

AI Solution Architect & Technical Product Manager (Consultant)

Remote, US

- Engineered **Indiadatacommons.org** by integrating Google Datacommons with Gemini and MCP, providing Google, MoSPI, and IIIT-B with automated policy-analysis tools that generated **\$7.5M USD in revenue**.
- Leading a diverse team of AI Engineers (ranging from freshers to 8+ years of experience) to build and deploy internal automations and custom GenAI solutions for the **Google Data Commons** team.
- Tech: **GenAI, Gemini, MCP, Google Data Commons, Google ADK, Python, JavaScript, GCP, Server Administration**

Indeep AI

July 2024 – Jun 2025

AI Solution Architect & Technical Product Manager (Consultant)

Remote, Spain

- Architected and launched **Typewiser.com**, an AI-driven platform that served as the foundational technology for **Dialogic**, securing a **2.5 million euro grant** from the European Innovation Council (EIC).
- Lead a team of 3 in orchestrating AI micro-services via Python FastAPI and ML Ops workflows, optimizing LLM performance and cost to scale from research prototype to a grant-funded commercial entity.
- Engineered the AI-driven assessment framework for the **ASTOUND project**, utilizing LLMs to evaluate human-machine intelligence benchmarks to advance the understanding of collective cognitive capabilities.
- Co-authored and published groundbreaking research in **Nature Human Behaviour**, validating the project's methodology and establishing a new scientific standard for comparing AI and human performance.
- Tech: **Python, FastAPI, Django, Machine Learning, LLMs, Agents, CI/CD, ML Ops**

Pitney Bowes

May 2022 – July 2024

Advisory Data Scientist

Noida, India

- Engineered an **LLM Agent-based business rules engine** to translate natural language into executable logic, automating rule-authoring and streamlining internal compliance operations.
- Architected high-precision ML models for **HS Code classification** and restricted item prediction on GCP, **saving in annual compliance costs** by reducing manual classification errors and audit overhead.
- Tech: **LLM Agents, Python, NLP, GCP, ML Ops, HS Code Classification, Predictive Modeling**

Precog @ IIIT-Delhi

Jun 2018 – May 2022

Research Scholar (AI & Privacy)

Delhi, India

- Directed a team of 20 to develop **AASMA**, a social media analysis engine adopted by **60+ Indian Government organizations**, and spearheaded cybersecurity policy for WISH Foundation to accelerate cross-entity data sharing.
- Developed **pyadel** for Infosys, a Python package for synthetic multi-table data generation with complex relationship constraints, now utilized by internal data science teams for high-fidelity modeling.

- Tech: **Python, Machine Learning, Synthetic Data, Blockchain, MongoDB, Cybersecurity Policy, Fullstack Webapps**

Amazon

Jan 2021 – July 2021

Applied Scientist Intern

Bangalore, India

- Engineered multilingual transformer models and transcription pipelines to enhance search relevance for Language of Preference (LOP) queries in Hindi, Tamil, and French.
- Achieved a **60-75x increase in query coverage** at 95% precision compared to legacy baselines, significantly improving search discovery for infrequent and broad multilingual queries.
- Tech: **AWS, Python, Machine Learning, NLP, Transformers, Multilingual Modeling, Transcription**

Google Summer of Code (mlpack)

May 2017 – July 2017

Open Source Contributor

Remote

- Migrated 345,000 test configurations from virtual machines to Docker containers, **achieving a 70% reduction in storage footprint** while optimizing the mlpack development environment.
- Engineered automated CI/CD pipelines using Jenkins for nightly and multi-frequency builds, utilizing a custom-built minimalist Linux image to accelerate development cycles.
- Tech: **Shell Scripting, Docker, Jenkins, Linux, CI/CD, Virtualization, mlpack**

Publications

- Manu Kumar Shetty, Aaloke Mozumdar, **Saurabh Gupta**, Lalit Gupta, Kapil Chaudhary, Vandana Roy, Suresh Kumar, Bhupinder Singh Kalra, Sanjeev Khanth PE, Anubha Gupta, Leveraging Explainable AI for Drug Efficacy and Mortality Analysis: Insights from an Observational Dataset, Accepted at Medinformatcs Journal, (2025). [Paper]
- Strachan JWA, Albergo D, Borghini G, Pansardi O, Scaliti E, **Gupta S**, Saxena K, Rufo A, Panzeri S, Manzi G, Graziano MSA, Becchio C (2024). Testing Theory of Mind in Large Language Models and Humans. accepted at Nature Human Behaviour.
- **S. Gupta**, A. B. Buduru and P. Kumaraguru, Differential Privacy: A privacy cloak for preserving utility in heterogeneous datasets, Accepted at CSI Transactions on ICT, (2022). [Paper]
- Chandra, M., Reddy, M., Sehgal, S., **Gupta, S.**, Buduru, A., and Kumaraguru, P. “A Virus Has No Religion”: Analyzing Islamophobia on Twitter During the COVID-19 Outbreak. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21), August 30-September 2, 2021, Virtual Event, Ireland. [Paper]
- **Gupta, S.**, Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020, November). Viable Threat on News Reading: Generating Biased News Using Natural Language Models. Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, 55–65. doi:10.18653/v1/2020.nlpccs-1.7. [Paper]
- **S. Gupta**, S. Bhambri, K. Dhingra, A. B. Buduru and P. Kumaraguru, “Multi-objective Reinforcement Learning based approach for User-Centric Power Optimization in Smart Home Environments,” 2020 IEEE International Conference on Smart Data Services (SMDS), 2020, pp. 89-96, doi: 10.1109/SMDS49396.2020.00018. [Paper]
- **S. Gupta**, A. B. Buduru and P. Kumaraguru, “imdpGAN: Generating Private and Specific Data with Generative Adversarial Networks,” 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2020, pp. 64-72, doi: 10.1109/TPS-ISA50397.2020.00019. [Paper]
- **S. Gupta**, A. K. Singh, A. B. Buduru and P. Kumaraguru, “Hashtags Are (Not) Judgemental: The Untold Story of Lok Sabha Elections 2019,” 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), 2020, pp. 216-220, doi: 10.1109/BigMM50055.2020.00038. [Paper]
- **Gupta S.**, Agarwal A., Vyalla, S., Buduru, A. and Kumaraguru, P. #IVoted to #IGotPwned: Studying Voter Privacy Leaks in Indian Lok Sabha Elections on Twitter. In 12th International Conference on Social Informatics (SocInfo) 2020. [Poster]

Education

Indraprastha Institute of Information Technology

Delhi, India

M.Tech. with Specialization in AI

Cluster Innovation Centre, University of Delhi

Delhi, India

B.Tech. in Information Technology and Mathematical Innovations

Technical Skills

- **Programming Languages:** Python (Advanced), JavaScript/TypeScript, SQL, HTML/CSS, Bash
- **AI/ML & GenAI:** Large Language Models (LLMs), Generative AI, Agentic Systems, Retrieval-Augmented Generation (RAG), Fine-tuning (LoRA, PEFT), Prompt Engineering, Function Calling, Tool Use, Multi-modal AI (Vision + Text), Synthetic Data Generation, Reinforcement Learning, Transformers, Diffusion Models, Evaluation Frameworks (LLM Evals, Guardrails)
- **LLM Frameworks & Tooling:** LangChain, LangGraph, LlamaIndex, HuggingFace, vLLM, OpenAI API, Gemini API, Anthropic Claude, AutoGen, CrewAI, Gradio, FastAPI
- **ML/DL Frameworks:** PyTorch, TensorFlow, scikit-learn, XGBoost, mlpack, SDV
- **MLOps & LLMOps:** Model Deployment, CI/CD for ML, Experiment Tracking, Model Monitoring, Prompt Versioning, Vector Databases (FAISS, Pinecone, Weaviate), Feature Stores, A/B Testing for Models, Cost Optimization, Latency Optimization
- **Cloud & AI Infrastructure:** AWS (ECS, EC2, S3, RDS, SQS, SES, Load Balancers, CodePipeline, CodeBuild, Secrets Manager), Google Cloud (Vertex AI, Agent Builder, Model Garden, Pipelines, RAG Engine), Docker, Kubernetes (basics), Serverless Architectures
- **Data & Backend Systems:** REST APIs, Microservices Architecture, FastAPI, Node.js, Express.js, Next.js, Django, MongoDB, MySQL, Elasticsearch, Data Pipelines
- **Developer Tools:** Git, Docker, VS Code, Linux CLI, Linters, Jupyter, Postman
- **Specialized Areas:** NLP, Multilingual Modeling, Knowledge Graphs, Semantic Search, AI Agents for Automation, Financial Modeling with AI, Privacy-preserving ML (Differential Privacy)
- **Operating Systems:** Linux (Advanced Server Administration), MacOS, Windows

Certifications

- Google Cloud Certified Professional Machine Learning Engineer
- Anthropic Claude Certified Professional