Sarthak Singh

asarthaks@gmail.com
Get Latest version

asarthaks
+49-15566271040
Schedule a Call

Professional Summary

4+ years of AI/ML experience across Industry (InsurTech, Healthcare, EdTech) and Research (Sony, Fraunhofer), specializing in LLM optimization, NLP, and AI-driven automation. Currently researching Controllable Text Generation (CTG) at Sony, focused on developing training-free, energy-based decoding frameworks using Langevin Dynamics to steer LLM outputs.
Proficient in building and deploying scalable AI solutions, including RAG frameworks, fine-tuned LLM systems, and high-throughput document processing pipelines. Skilled at bridging the gap between complex research (SGLD, MCMC sampling) and business requirements to drive efficiency. In my free time, I actively engineer Multi-Agent Systems and agentic workflows to automate personal productivity and explore the latest in autonomous LLM orchestration.

Education

- **Universität Stuttgart** — Stuttgart, Germany
  M.Sc. Computational Linguistics — *October. 2023 – Present*
  - **Master's Thesis**: Controllable Text Generation Using Langevin Dynamics (In collaboration with **Sony Research**)
  - **Courses**: Introduction to Deep Learning, Advanced Deep Learning, Reinforcement Learning, Knowledge Graphs, Foundation Models, Machine Perception and Learning, Speech Technology, Machine Learning, Modelling Semantic Plausability
  - **Cumulative GPA**: 1.2
- **Guru Gobind Singh Indraprastha University** — New Delhi, India
  B.Tech. Computer Science and Engineering — *July. 2015 – June. 2019*
  - Courses: Machine Learning, Artificial Intelligence, Applied Linear Algebra, Data Structures & Algorithms

Skills Summary

- **Languages**: Python, C, SQL, Unix & Shell Scripting, Latex, JavaScript
- **Tools**: SLURM, Stune, Numpy, Pandas, PyTorch, Transformers, Langchain, Scikit-Learn, Matplotlib, XGBoost, OpenCV, Huggingface, Git, Pytest, NLTK, SpaCy, Flask, Kubernetes, Docker, GIT, Airflow, AWS, GCP, MLC, TVM, Distilable, Django, PostGreSQL, OpenClaw, Ollama
- **MLOPs**: Airbyte, Great Expectations, Weights & Biases, MLFlow, DVC, Sagemaker, Lambda, ECR, ECS, Docker, FastAPI, Streamlit, Gradio, AirFlow, ONNX Runtime, AutoGEN, Deepspeed, Megatron LM, VLLM, Tailscale, Podman, KubeFlow, Kafka, Apache, Gunicorn
- **Technical Skills**: Energy-Based Models (EBMs), Langevin Dynamics (SGLD), Markov Chain Monte Carlo (MCMC), Inference-time Guidance, Zero-shot Steering, Bayesian Inference, Posterior Sampling, Large Language Modeling and Pre-training, Supervised Fine Tuning, DPO, PPO, Deep Learning, Machine Learning, Deep Learning, Natural Language Processing, Model Deployment, Model APIs, Multi Node Training
- **Soft Skills**: Leadership, Planning and Management, Writing, Extempore, Time Management

Experience

- **Sony Research** — Stuttgart, Germany
  AI Research Intern & Master's Thesis Student — *June 2025 - Present*
  - **Energy-Based Controllable Generation**: Developing a training-free framework for steering LLMs (Llama 3 8B) toward global constraints (sentiment, toxicity) using Energy-Based Models (EBMs), enabling zero-shot attribute control without model fine-tuning.
  - **Langevin Dynamics Optimization**: Implementing Stochastic Gradient Langevin Dynamics (SGLD) to navigate discrete token spaces by leveraging gradients from continuous embedding spaces, improving the reliability of non-autoregressive decoding.
  - **Sampling Fidelity & Bias Mitigation**: Integrated a Metropolis-Hastings (MH) correction step into the Langevin sampler to mitigate discretization bias, ensuring the model explores the true posterior distribution rather than converging to local maxima (noisy gradient descent).

- **Discrete Langevin Sampler (DLS)**: Formulated and tested a transition kernel for discrete variables to reconcile the non-differentiable nature of text with gradient-based proposals, achieving a better trade-off between control adherence and linguistic diversity.
- **Advanced Evaluation Benchmarking**: Quantifying performance using Perplexity (PPL) for fluency, MAUVE for distributional quality, and Self-BLEU for diversity, while monitoring efficiency via Number of Function Evaluations (NFE).

- **University of Stuttgart - Institut für Maschinelle Sprachverarbeitung** Stuttgart, Germany
  Research Assistant - Graph Query Parsing *March 2025 - May 2025*
  - **Graph-based Query Parsing**: Developed algorithms to traverse a domain-specific graph to resolve user queries by identifying the closest matching nodes using graph traversal techniques.
  - **LLM Integration**: Self-hosted and integrated a **DeepSeek 14B** model to enhance natural language understanding, enabling the chatbot to better interpret and match free-form queries to graph nodes.
  - **Dialogue Breakdown Detection**: Researched and implemented techniques for detecting dialogue breakdowns to maintain coherent and context-aware conversations in the chatbot system.

- **University of Stuttgart - Institut für Maschinelle Sprachverarbeitung** Stuttgart, Germany
  Teaching Assistant - Introduction to Deep Learning *January 2025 - February 2025*
  - **PyTorch & Neural Networks**: Conducted tutorials on PyTorch fundamentals, covering preprocessing, datasets, data loaders, and training neural networks (**CNNs, RNNs, Transformers**).
  - **Transformers & NLP**: Taught students how to use the Hugging Face Transformers library for tasks like token **classification, sentiment analysis**, and **text generation** using decoder-only models.
  - **Exercises & Exam Preparation**: Designed and presented weekly exercises and solutions. Created a questionnaire with PyTorch-based questions to help students prepare for exams.
  - **Live Coding & Final Project**: Held **live coding sessions** to assist students with their final project, where they trained models to solve real-world tasks.

- **University of Stuttgart - Institut für Visualisierung und Interaktive Systeme** Stuttgart, Germany
  Research Assistant - Human Motion Forecasting *November 2024 - May 2025*
  - **Human Motion Prediction**: Worked on predicting human motion using **3D data** from AR-VR devices like **HOT3D**, leveraging methods such as HOIMotion.
  - **Intent & User Classification**: Analyzed **eye-gaze data** using **DenseNets** to predict user intent and classify users based on their motion patterns.
  - **Pose Estimation & GCNs**: Developed models for **joint position prediction** using motion and gaze data, employing **Pose Residual GCN** and **Fusion GCN** architectures.

- **Fraunhofer-Gesellschaft** Stuttgart, Germany
  Working Student - Generative AI *June 2024 - September 2024*
  - **LLM Evaluation**: Researched state-of-the-art LLM evaluation benchmarks and implemented **MMLU Pro** and **MultiWOZ** for in-house model assessment.
  - **Retrieval-Augmented Generation**: Optimized Retrieval-Augmented Generation (RAG) performance by implementing state-of-the-art evaluation benchmarks (MMLU Pro, MultiWOZ) and refining the retrieval-based LLM orchestration layer.

- **Gigaforce Private Limited** Noida, India
  Senior Software Engineer - AI/ML *October 2022 - September 2023*
  - **SubroGPT**: Designed and curated a Custom Dataset using **Falcon 40B** for Subrogation Domain. **Instruct fine-tuned** an open source LLM with 7B parameters using **PEFT** and **LoRA** to perform **transfer learning**. Applied it to several downstream tasks like **email generation, chain of thought answering, and context understanding**.
  - **QA Retrieval Agent**: Engineered a robust QA Retrieval agent by leveraging **Langchain, FastChat5B, and ChromaDB Vector Store**. Analyzed accident descriptions and traffic laws to identify law infringements, streamlining liability determination alongside SubroGPT. Increased claim closure rate by 500%.
  - **Data Extraction from Police Report**: Utilized **OpenCV** and **Tesseract** to identify and extract structured data from Police Reports. This automated data extraction resulted in a decrease of claim processing time by 1 hour/report.
  - **Subro Potential Application**: Created a production-ready application that could successfully analyze a claim and determine its subrogation potential. Used **TFiDF, NER, BERT Embeddings and sentence similarity** to understand the data, gather features and trained on historical data. **Reduced the claim processing time from 100 claims/day to 10,000 claims/hour** while **increasing per claim revenue by 40%**. Used **Dockers** to deploy the application and **Airflow** for scheduling and monitoring. Daily processing 1M+ documents to predict subrogation potential
  - **AI Enabled Claim Document Parser**: Developed an Algorithm to identify patterns and extract data from an unseen pdf while maintaining its structure. Went from POC to deployment in 1 month. This allowed streamlined text extraction resulting in significant time reduction and huge savings for clients. 250K+ documents processed till date

- **Impact**: Successfully took several AI/ML projects from proof-of-concept to production. Developed and applied advanced algorithms in the domain of **subrogation opportunity identification, liability analysis, and text extraction** leading to increased subrogation efficiency and revenue for clients. Identified need for Airflow, **WandB** incorporation, set up and managed airflow and WandB to automate, schedule, monitor, and debug AI applications and artifacts

- **Navia Life Care** Gurgaon, India
  Machine Learning Engineer *August 2019 - September 2022*
  - **Handwriting Annotation and Recognition System**: Built an end-to-end system for collecting and labeling doctor's handwritten prescriptions, Leveraged collected data to enhance **handwriting and character recognition on doctor's prescriptions** through the utilization of **CNN, RNN, and CTC techniques** achieving a **90% accuracy rate in offline recognition**
  - **Clinical Decision Support System**: Implemented personalized **recommendation engine** for 20k+ doctors based on patient's information, symptom details and usage data
  - **ETL Pipelines and Reporting System**: Led the development of an ETL pipeline using PostgreSQL and Django, in conjunction with a reporting system featuring HTML, CSS, and JavaScript visualizations. Automated data pipelines and reporting for extensive doctor and patient records, alongside the development of APIs for seamless integration.

Publications & Presentations

S. Singh Stuttgart, Germany
Poster Presentation: Immersive text-to-speech systems *[Current Topics in Speech Technology, IMS], February, 2025*
- Portfolio: Click to view more details

S. Singh Stuttgart, Germany
Poster Presentation: Direct Preference Optimization (DPO) *[Reseach Seminar, IMS], January, 2025*
- Portfolio: Click to view more details

Achievements

- Recipient of the **Deutschlandstipendium** 2025 Scholarship for Academic Excellence
- Kaggle Expert
- Achieved top 40 ranking out of 6000 students in the ZS Data Science Challenge 2018.