Ashish Jha

ML Research Engineer

Goettingen, Germany | EU Work Permit Available | Available from June 2026
+49 155 602 47407 | ashish.jhav1@hotmail.com
LinkedIn | GitHub | Google Scholar

Professional Summary

ML Research Engineer with experience in LLM post-training (SFT, RLHF, DPO), distributed GPU training, and production ML pipelines. Published in IEEE (Q1) and NeurIPS 2025. Built open-source PyTorch tooling (GRAFT, PyPI). Based in Goettingen, Germany EU work authorisation, available immediately.

Work Experience

Research Engineer - Data Science & ML Jun 2025 – Apr 2026
Applied AI Institute

- Designed and ran end-to-end LLM post-training pipelines (SFT → RLHF → DPO) for code generation tasks on LLaMA and Mistral models; systematically benchmarked alignment strategies against HumanEval baselines.
- Experimented with alternative divergence-based reward objectives beyond standard DPO/KL; quantified a ~2.7x compute overhead with marginal alignment gains, informing production training strategy decisions.
- Built custom evaluation harnesses measuring pass@k and resolve rate across model versions and training configurations; results directly informed model release decisions.
- Generated and curated synthetic training datasets to improve model quality across task configurations and training runs.
- Ran distributed training and evaluation on Slurm clusters (H100, H200 GPUs); deployed model on GCP served via REST APIs.

ML Engineer(MLOps) Jan 2025 – Mar 2025
Sartorius AG, Goettingen

- Delivered an end-to-end ML pipeline – data preprocessing, feature engineering, training, and evaluation – fully to production; adopted by the operations team and replacing all prior manual workflows.
- Benchmarked 5+ model architectures (LSTM and baselines) on Azure ML across 200+ features and ~25K records with 20–25% missing data from non-operational factory sites.
- Built 2 automated data ingestion pipelines and automated ETL workflows on Azure Data Factory using Apache Spark across multiple factory regions.
- Automated business reporting in Python, fully replacing manual analyst workflows – adopted across the operations team.

Doctoral Researcher – Data Science & ML Oct 2021 – Dec 2024
Skolkovo Institute of Science and Technology (Skoltech)

- Developed GRAFT – a PyTorch library for intelligent training-data selection using gradient information and feature decomposition – reducing training computation by up to 50% while maintaining benchmark accuracy. Published at NeurIPS 2025 (workshop); released on PyPI as open-source.
- Built end-to-end training and evaluation pipelines in PyTorch and Hugging Face Transformers covering data loading, model training, hyperparameter tuning, and reproducible benchmarking.
- Ran large-scale distributed training experiments across 4 GPUs in parallel on Slurm clusters.
- Researched and implemented model compression techniques including structured pruning, quantisation, and knowledge distillation for efficient deployment.
- Tracked and compared experiments using Weights & Biases (W&B), enabling data-driven model release decisions across training runs.

ML Research Engineer (Joint Research Project) Feb 2022 – Oct 2022
Huawei Technologies & Skoltech

- Researched and developed novel autoencoder-based architectures in PyTorch for end-to-end image and video compression, targeting improved rate–distortion performance over conventional codecs.
- Designed and implemented training pipelines for variational and learned compression models, including custom loss functions balancing reconstruction quality and bitrate.

- Built evaluation frameworks to benchmark compression models against standard codecs (JPEG, H.264/H.265) on PSNR, MS-SSIM, and bitrate metrics across multiple datasets.
- Collaborated directly with Huawei's research team on-site, translating experimental findings into actionable recommendations for production compression pipelines.

Software Engineer (Python) Aug 2019 – Jul 2021
Ruijie Networks, St Petersburg

- Built Python-based automation tools and internal CLI libraries for large-scale data collection, log parsing, and anomaly detection across enterprise network infrastructure.
- Developed data extraction and processing pipelines (REST APIs, JSON/CSV ingestion) to normalise and analyse high-volume system data for pattern identification and reporting.
- Designed automated testing and validation workflows using HTTP methods with `requests`/`urllib`, reducing manual effort by ~40%.
- Built reporting and data aggregation scripts in Python (Pandas, JSON processing) to produce structured analysis reports for engineering and management stakeholders.

Technical Executive – Data & Infrastructure Aug 2016 – Feb 2019
CG Communications Ltd., Nepal

- Designed and maintained high-availability cloud infrastructure for a payment platform; automated deployment workflows in Bash and Python, reducing downtime by over 50%.
- Oversaw data infrastructure and operational systems across the organisation, supporting scalable and reliable service delivery.

Education

Doctoral Researcher in Computational Data Science 2021 – 2025
Skolkovo Institute of Science and Technology (Skoltech)
Research focus: Efficient deep learning, model optimisation, data-efficient training methods

M.Sc. in Computer Science 2019 – 2021
Saint Petersburg Electrotechnical University (LETI)

B.E. in Electronics & Communication Engineering 2010 – 2014
Visvesvaraya Technological University (VTU)

Publications

Lightweight Attribute Localizing Models for Pedestrian Attribute Recognition 2026
IEEE Intelligent Systems **[Q1 Journal]** | Link

SAGE: Streaming, Agreement-driven Gradient Sketches for Representative Subset Selection 2025
NeurIPS 2025, Workshop | Link

Skills

Programming	Python (expert), SQL, C++ (familiar)
ML Frameworks	PyTorch, Hugging Face Transformers, Weights & Biases
LLM Post-Training	SFT, RLHF, DPO, preference optimisation, reward modelling, QLoRA, PEFT/LoRA
Model Optimisation	Quantisation, structured pruning, knowledge distillation, efficient fine-tuning
Data & Synthesis	Synthetic data generation, Apache Spark, Pandas, large-scale data pipelines
Distributed Training	Slurm clusters, multi-GPU training (H100, H200), experiment reproducibility
Deployment & MLOps	Docker, FastAPI, CI/CD (GitHub Actions), Git, Azure ML, Azure Data Factory
Evaluation	HumanEval, SWE-Bench, pass@k, custom harness design, benchmark analysis

Certifications

- Databricks Machine Learning Professional Certification

Languages

English (Fluent) | Hindi (Native) | Nepali (Native) | Maithili (Native) | Russian (Limited Working) | German (Basic, A2)