debanjan.basu.ds@gmail.com
Berlin, Germany
github.com/d3banjan
d3banjan.github.io/engineering

SKILLS

Agent / Eval
Arize Phoenix · butterflow · LangChain · LlamaIndex · RAG · vector DBs · token economics

Distributed
Dask · Celery · Django · PostgreSQL · S3/MinIO · Docker · Kubernetes

ML / Research
PyTorch · HuggingFace · LoRA/PEFT · GPTQ · CUDA · Lean 4

EDUCATION

B.S.–M.S. Physics
IISER Kolkata, 2007–2012

Doctoral research (not completed)
TU Clausthal, 2012–2016

LANGUAGES

English (fluent) · German (B1)
Hindi (native) · Bengali (native)

Debanjan Basu

SENIOR ML ENGINEER · DISTRIBUTED SYSTEMS · LLM OBSERVABILITY

Six years production ML engineering at Nexern (Berlin): LLM agent observability (Arize Phoenix), distributed pipelines (Dask/Celery, 5–10× speedups), GenAI agent deployment. Built **butterflow** — declarative agent eval framework with token caching between runs; **falcon** gates unsafe ML deserialization at the type-checker level via Lean 4-verified typestubs.

SELECTED PROJECTS

butterflow

github.com/d3banjan/butterflow

- CLI framework: declarative agent flow definitions, token caching between test runs, evals and cost optimization unified
- Same execution trace, cache aggressively, measure quality simultaneously — user-flow testing and token cost as one problem

falcon

github.com/d3banjan/falcon

- Python typestubs tracking payload annotations by source to gate unsafe ML deserialization (pickle, HDF5) at the type-checker
- Born from huntr.com CVE research on serialization-route vulnerabilities in GenAI platforms; Lean 4 soundness proofs included

EXPERIENCE

Nexern GmbH
Senior ML & Data Engineer · Aug 2020–Jul 2026 · Berlin

- Built LLM agent observability with Arize Phoenix; trajectory analysis for debugging agent failures in production
- Achieved 5–10× pipeline speedups via Dask distributed processing, Celery task queues, and vectorised operations
- Operational exposure to KV-cache memory pressure and throughput–latency tradeoffs on long-context agent runs
- Led greenfield GenAI agent projects: LLM-based prediction pipelines and automated web data extraction
- Django data platform: large-scale JSON ingestion, web crawlers, Docker deployment, GitLab CI/CD

KUGU Home GmbH
Data Scientist · May 2018–Jul 2020

- Physics-informed heat models and time series forecasting with TensorFlow/XGBoost
- IoT pipeline for hundreds of devices — online changepoint detection, OpenStack/Ansible deployment

TU Clausthal
Doctoral Researcher · 2012–2016

- Extended MD codebase (Fortran/C); published in *Physica Status Solidi A* (DOI: 10.1002/pssa.201532488)

RESEARCH

Independent, Oct 2025–present. 5 preprints across MoE quantization, LoRA-DPO alignment geometry, RoPE provenance, symmetry survey, and compression methodology. 74 Lean 4 theorems (Mathlib).
d3banjan.github.io/research