Debanjan Basu

Research Engineer / ML Systems Engineer · LLM Infrastructure · Observability

debanjan.basu.ds@gmail.com · Berlin, Germany · github.com/d3banjan · huggingface.co/d3banjan · d3banjan.github.io

PROFILE

ML systems engineer with independent research bridging post-training methods, formal verification, and empirical methodology for safety-relevant claims. LoRA-DPO scaling work on Pythia 70M–1B documents geometry–behaviour decoupling: γ-overlap doesn't predict reward margin, suggesting behavioral interventions don't necessarily reorganize underlying representations — a structural finding with implications for evaluating alignment techniques. Lean 4 invariants formalize the empirical scaling (74 theorems across two papers, all in Mathlib). Pre-registered adversarial validation methodology — trap-cell design, kill-fast sequential testing — applied across MoE compression rungs. Six years production ML engineering at Nexern: LLM agent observability with Arize Phoenix, distributed pipelines, deployment infrastructure. Physics background (IISER Kolkata BS-MS; doctoral research under Peter Blöchl at TU Clausthal).

EXPERIENCE

Senior ML & Data Engineer Aug 2020 – Jul 2026

Nexern GmbH, 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
- Built and maintained Django data platform; large-scale JSON ingestion and web crawlers
- Deployed agents with Docker on VPS; S3/MinIO storage; GitLab CI/CD

Data Scientist May 2018 – Jul 2020

KUGU Home GmbH, Berlin
- Developed physics-informed heat models (Fourier equation) and time series forecasting with TensorFlow/XGBoost
- Built online changepoint detection for anomaly detection in IoT boiler systems
- Co-developed IoT pipeline for hundreds of devices; deployed to OpenStack via Ansible
- Ensured GDPR compliance in data handling workflows

Doctoral Researcher (Physics) Aug 2012 – Oct 2016

TU Clausthal (Institute for Theoretical Physics) & University of Göttingen
Supervisor: Prof. Peter E. Blöchl
- Investigated phonon dynamics and thermoelectric transport via classical molecular dynamics
- Extended MD codebase (Fortran/C): force constant extraction, phonon bandstructure calculation, thermal transport properties
- Published in *Physica Status Solidi A* (DOI: 10.1002/pssa.201532488)
- Tutored ab-initio electronic structure methods

SKILLS

ML Systems / Infrastructure
Python, Django, Docker, Kubernetes, AWS (S3, EC2), GitLab CI/CD, Ansible, PostgreSQL, S3/MinIO
Distributed Data
Dask, Celery, vectorised ops, large-scale JSON ingestion, web crawlers
LLM Observability / Evals
Arize Phoenix, trajectory analysis, agent debugging, TensorFlow, scikit-learn, XGBoost
Research & Formal Methods
Lean 4 / Mathlib, PyTorch, HuggingFace Transformers/ Accelerate, LoRA/PEFT, quantisation (GPTQ, INT4/INT8), SVD/spectral methods, CUDA, Fortran, C, Rust

RESEARCH

Independent research, Oct 2025 – present

Phase-Collapse Defragmentation: A Moment-Ratio Framework for 1-Bit KV-Cache Quantization in MoE Transformers 2026

preprint microsite with Lean source

- Learned orthogonal rotation lifts moment-ratio cosine β from SRHT floor (0.80) to 0.92–0.97 across four architectures (Gemma-4 e2b/e4b/26B-MoE, DeepSeek-MoE-16B, OLMoE-1B-7B)
- Discovered Stiefel frustration: MoE expert banks resist single-rotation alignment at $\beta \approx 0.83$
- 74 Lean 4 theorems proved (Mathlib); includes convergence bounds and quadratic last-mile hardness

Low Stable-Rank Structure in LoRA-DPO Adapters on Pythia 70M–1B: Empirical Scaling and Formal Invariants 2026

preprint microsite with Lean source

- Stable rank ≈ 3.6 floor across 4× width scaling under fixed LoRA-DPO recipe
- Geometry–behaviour decoupling: γ-overlap does not predict reward margin
- Lean-verified: `stableRank_smul_invariant`, `rsLoraUpdate_frob_bounded`
- All adapter checkpoints released on HuggingFace (≈1.9 GB)

Symmetry Survey for Verified Neural Compilation 2026

in progress

- Catalog of transformer weight symmetries (RoPE-commuting rotations, sign/phase gauges, parabolic stabilizers) with Lean 4 companion

The Compression Falsification Ladder compression-ladder-paper.pages.dev

- Pre-registered empirical methodology for honest compression research: SHA-locked protocols, trap-cell adversarial validation, τ-hardened random baselines, kill-fast sequential design
- Applied to 10+ compression rungs on OLMoE-1B-7B; 7 clean kills, 1 deepen-strict result (per-channel INT4), 1 impact-rung in flight

PUBLICATIONS

Peer-reviewed

Basu, D. & Blöchl, P.E. (2016). Phonon dynamics and thermoelectric transport in thermoelectric materials. *Physica Status Solidi A*. DOI: 10.1002/pssa.201532488

Preprints and research artifacts (2026)

Basu, D. Phase-Collapse Defragmentation: A Moment-Ratio Framework for 1-Bit KV-Cache Quantization in MoE Transformers. Microsite with Lean source.

Basu, D. Low Stable-Rank Structure in LoRA-DPO Adapters on Pythia 70M–1B: Empirical Scaling and Formal Invariants. Microsite with Lean source.

EDUCATION

Doctoral research (not completed) 2012 – 2016

University of Göttingen / TU Clausthal

B.S.–M.S. in Physics 2007 – 2012

Indian Institute of Science Education and Research (IISER) Kolkata

LANGUAGES

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