

Rahul Krishnan

M.Sc. Data Science, ML Systems and LLM Inference
Trier, Germany

+49 15510269942, rahulkrishnan1105@gmail.com, GitHub, LinkedIn, Website

PROFILE

M.Sc. Data Science candidate focused on ML systems, efficient LLM inference, and reliable AI. First author of *A JoLT for the KV Cache*, which achieves near-lossless 2x-3x KV-cache compression on Mistral-7B and LLaMA-2-13B. I have also conducted reinforcement-learning research on reward design, calibration, and abstention. Before my master's, I spent more than four years at SAP Labs India building reliable production systems at scale. I combine empirical ML research with production software engineering and enjoy turning technically difficult ideas into working systems.

SELECTED RESEARCH

- MSc Thesis: Near-Lossless KV-Cache Compression

Mar 2026 - Present

- Developed a runtime KV-cache compression method combining partial Tucker decomposition, low-bit residual coding, and constrained per-layer allocation; released as the first-author preprint *A JoLT for the KV Cache* and now being extended for an ICLR 2027 submission.
- Achieved near-lossless 2x-3x compression on Mistral-7B and LLaMA-2-13B: perplexity changed by approximately 0.1-0.4%, reasoning accuracy remained within baseline confidence intervals, and needle-in-a-haystack retrieval was preserved through 8K context.
- Currently developing a custom fused-attention kernel in Triton that operates directly on compressed KV factors; an intermediate reconstruction-cache implementation reduced decode overhead from a 75x-100x naive projection to approximately 3.7x the uncompressed DynamicCache baseline.

- Asymmetric RL Rewards for LLM Calibration and Abstention (FEAR)

Mar - Jun 2026

- Built an end-to-end GRPO RLVR training and calibration-evaluation stack on Qwen 2.5 1.5B and 3B, covering seven reward conditions and six math and reasoning benchmarks.
- Modified TRL training internals to compare z-score, RLOO, and signed-reward advantage estimators; integrated in-process vLLM rollouts, LoRA adapters, and calibration metrics including ECE, Brier score, AUROC, and coverage-risk curves.
- Proved that within-group z-score normalization makes two-outcome GRPO rewards invariant to reward magnitude, reducing asymmetric rewards to an abstention-threshold effect; confirmed experimentally at both model scales, with frontier-condition accuracy matching within 0.004.
- Ran training and evaluation across A100, V100, and 2080 Ti systems, diagnosing vLLM/PEFT memory co-residency, mixed-precision compatibility, and GPU-memory stability failures.

EXPERIENCE

- Student Research Assistant (studentische Hilfskraft)

Apr 2026 - Present

Universität Trier, Department of Mathematics

Trier, Germany

- Continue the KV-cache compression and tensor-optimization research under Prof. Dr. Volker Schulz following the thesis study, including the Triton kernel implementation and paper extension.
- Previously graded weekly Numerical Optimization exercises on line-search and trust-region methods, KKT conditions, and numerical linear algebra, providing written feedback across the course cohort.

- Software Developer, SDE I to SDE II

Jul 2020 - Sep 2024

SAP Labs India Pvt. Ltd

Bengaluru, India

- Owned one of three core modules of SAP's Document Consistency Cockpit end-to-end, designing its schema, extensibility model, retry semantics, and idempotency layer; load-tested it over 10M-100M document-processing cycles and deployed it for Spain and India.
- Designed a declarative XML content-matching framework that reduced new-country onboarding from approximately 10 days of custom development to 3 days of configuration; adopted by Spain, India, and Chile teams.
- Replaced approximately 100 paginated API calls with one bulk-retrieval request on the Spain reporting path, reducing report-generation time by approximately 30%.
- Initiated and led a Turkey error-handling overhaul based on two years of incident data; incident volume fell approximately 88% and customer support ratings increased from 2-3 to 4-5 on a five-point scale.
- Promoted through successive engineering levels to SDE II, received an SAP Stellar Award, and was selected as the only person from my area for SAP Labs India's fast-track architect career-development program.

ADDITIONAL RESEARCH

- Reflective CoT Faithfulness Study (Think Twice)

Apr - Jun 2025

- Co-built a controlled evaluation of chain-of-thought faithfulness across Gemini 2.5 Flash Lite and LLaMA 3.2 3B on 650 BIG-Bench Hard examples; reflective CoT reduced bias susceptibility by approximately 7 percentage points but did not eliminate systematic post-hoc rationalization.

- Risk-Averse Optimization with CVaR

Oct 2025 - Jan 2026

- Applied Rockafellar-Uryasev CVaR optimization to nonlinear optimization, robust SVM training, and model-predictive control; a risk-aware fermentation controller reduced constraint violations from 100% to 0% under severe model mismatch.

- Neuro-Symbolic Narrative Similarity

Nov 2025 - Jan 2026

- Built a heterogeneous GNN over LLM-extracted narrative graphs and fused it with text embeddings, reaching 72.5% pairwise accuracy; principal-angle analysis showed the graph and text channels occupied near-orthogonal subspaces.

PUBLICATIONS

- | | |
|--|------|
| - A JoLT for the KV Cache– arXiv:2607.12550, R. Krishnan and V. Schulz | 2026 |
| - Glaucoma Detection from Retinal Fundus Images– IEEE ICCSP 2020 | 2020 |

EDUCATION

- | | |
|--|--------------------------------|
| - Master of Science in Data Science | Oct 2024 - Sep 2026 (Expected) |
| Universität Trier, Germany | Grade: 1.5 (German Scale) |
| - Bachelor of Technology in Computer Science and Engineering | 2016 - 2020 |
| Amrita Vishwa Vidyapeetham, India | Grade: 1.37 (Top 10 in cohort) |

SKILLS

Core: Python, PyTorch, JAX, Triton, Transformers, Hugging Face, vLLM, TRL, LoRA / PEFT
ML Systems: LLM inference, KV-cache optimization, GPU kernel development, tensor decomposition, quantization, mixed precision, performance debugging, model evaluation
Reinforcement Learning: GRPO, RLVR, reward design, advantage estimation, calibration, selective prediction
Engineering: Git, Linux, Docker, CI/CD, GitHub Actions, FastAPI, Weights & Biases, GCP, Apache Spark
Tools: Claude Code, Codex; **Additional:** Java, SQL, Rust, R; **Languages:** English (C2), German (B1)