Keshav Krishna

New York City, US | +1 (479) 685 5463 | kk6081@nyu.edu
GitHub | Website | LinkedIn | Google Scholar

Product engineer with three years shipping ML and NLP systems that consumers touched every day, most recently in an app serving 100M+ users. I own features from an ambiguous conversation to production, and I already work through LLM agents daily — writing the specs, evals, and verification gates rather than trusting output on faith.

Experience

Moore Capital Management — Risk Technology Intern May 2026 – Present
Risk analytics tooling and dashboards for internal users *New York City*

- Building internal risk dashboards and reporting tools used by risk and operations stakeholders; own features end to end, from a verbal request to something in production days later.
- Python data workflows feeding those surfaces, with the interface layer in HTML/CSS; iterating directly on feedback from the people who use the tools daily rather than through a spec handoff.

JioSaavn — Software Engineer, Data Science Jul 2023 – Jul 2025
Production ML and NLP in a consumer app serving 100M+ users *Mumbai, India*

- Built recommendation and personalization systems on embeddings and vector retrieval (Spark, Hive, Redis, Qdrant), driving **1M impressions, 800K clicks, and 500K streams per day** in production.
- Shipped a LLaMA3-8B-Instruct conversational playlist system doing intent understanding and retrieval-driven song selection — an LLM feature in front of real users, deployed on Docker and Kubernetes, with all the latency, cost, and failure-mode work that implies.
- Ran a model bake-off across SetFit, HingRoBERTa, mURIL, and LLaMA3-8B-Instruct for multilingual playlist moderation, reaching 85% accuracy against a manually labeled evaluation set; the model got picked on measured quality, not on vibes.
- Worked across product, editorial, and infrastructure teams in a small data-science group where the same person wrote the model, the pipeline, and the deployment — including the unglamorous parts nobody scopes.

GE Healthcare — Software Intern, Medical Computer Vision May 2022 – Jul 2022
Detection and de-identification pipelines for clinical imaging *Bangalore, India*

- Built YOLO-based medical-image de-identification and RGB/IR people-detection pipelines over 20K+ images.

Products and Agent Infrastructure I've Built

Job Application Command Center — React app on the Anthropic API *2026*

- React dashboard that pulls live job listings and generates tailored cover letters and STAR-format resume bullets per role, with kanban pipeline tracking across applications.
- Structured-JSON prompting with client-side parsing, validation, and error handling, so a malformed model response degrades gracefully instead of breaking the view.

Self-hosted agent stack — always-on Linux VPS *2026*

- Run an autonomous job-application agent with MCP tool integration and human-in-the-loop verification gates at the steps where a wrong action is expensive.
- Swapped model backends and hardened browser automation after real failures in production use — the unglamorous part of making an agent something you'd actually let run unattended.

FlashAttention-2 — from-scratch Triton implementation *2026*

- Forward and backward kernels with online softmax, tiling, causal masking, and activation recomputation; cut attention memory from $O(N^2)$ to $O(N)$, enabling 65K-token sequences.
- Profiled Transformer models up to 2.7B parameters with NVIDIA Nsight Systems and PyTorch memory snapshots, benchmarked against vanilla PyTorch and `torch.compile` baselines.

Transformer language model from scratch *2026*

- 17M-parameter model in PyTorch built from low-level components (BPE tokenization, RoPE, RMSNorm, SwiGLU, causal attention, AdamW, cosine LR scheduling), with systematic ablations across pre- vs. post-norm, RoPE vs. NoPE, and SwiGLU vs. SiLU.

Research and Publications

Phase-Adaptive Generation: Efficient Decoding for Diffusion Language Models *2026*

- NYU project under Prof. Greg Durrett. Analyzed generation dynamics of Dream/LLaDA-style diffusion LMs via per-token refinement traces, revealing interpretable phase structure: template spans stabilize immediately while arithmetic and answer-finalization spans demand substantially more refinement.
- Trained Transformer predictors over 138K block-level control tuples to forecast per-block refinement budgets; integrated into the LLaDA decoding loop, matching AdaBlock's 89.5% GSM8K accuracy with ~21% fewer forward evaluations. [Report]

Alignment and Reasoning RL for Mathematical LLMs *Spring 2026*

- End-to-end post-training pipeline: zero-shot evaluation, SFT, GRPO with verified rewards, masked losses, entropy tracking, and policy-gradient updates. Fine-tuned Qwen2.5-Math-1.5B on chain-of-thought data, improving MATH accuracy from 56.4% to 60.4%.

ML-Driven Program Phase Prediction for Dynamic Hardware Resource Optimization *Jan 2026 – Present*

- Independent study with Prof. Mohamed Zahran, NYU Courant. Unsupervised clustering over temporal traces of multicore workloads defines interpretable pressure states, distilled into compact predictors with sub-KB model state.

Publications *2024 – 2025*

- *Advancements in Cache Management: A Review of Machine Learning Innovations* — Frontiers in AI, 2025. Single-author, 10+ citations; reviewed 40+ papers on ML for cache management.
- *LLC Intra-set Write Balancing for Non-Volatile Memory Caches* — arXiv, 2024. Bachelor's thesis research.

Education

New York University, Courant Institute — MS, Computer Science Aug 2025 – May 2027
GPA 3.95/4.0; Machine Learning, Building LLM Reasoners, Honors Algorithms *New York City*

Indian Institute of Technology, Ropar — BTech, Computer Science and Engineering Aug 2019 – Jun 2023
GPA 8.74/10; Merit Scholarship for top academic performance *Ropar, India*

Technical Skills

Product / full-stack: React, JavaScript, HTML/CSS, FastAPI, Postgres, Redis, Docker, Kubernetes, Git, Linux

LLMs in production: Anthropic and OpenAI APIs, agents and tool use (MCP), evals and benchmarking, context management, RAG and vector search (Qdrant), SFT/GRPO/RLHF, PyTorch, Hugging Face, tokenizers, vLLM

Data / infrastructure: Python, C/C++, Spark, Hive, Airflow, Triton kernel development, NVIDIA Nsight Systems, Slurm/HPC, reproducible pipelines