

Kartikeya Agarwal | Machine Learning Engineer 2

Bengaluru, India 560103

☎ +91-9718502020 • ✉ kartikeya72001@gmail.com • in [kartikeya72001](#)
🌐 [kartikeya72001](#)

Professional Summary

Applied Machine Learning Engineer with 3+ years of experience architecting large-scale production ML systems, GenAI pipelines, and predictive risk engines at Navi Technologies. Engineered an automated 10-stage Databricks LLM evaluation DAG achieving 87.7% template accuracy and 99.0% precision on Tier 1 financial entities (-74.3% error reduction). Formulated a Cobb-Douglas template prioritization function and greedy submodular selection algorithm, expanding weekly template intake 20x (from 5k to 105k+ templates) and boosting SMS population coverage from 4.5% to 91.3% across 626M weekly messages. Architected a Flat-Array Compressed Sparse Row (CSR) Trie engine scaling across 8.39M templates on 100-node Databricks Spark clusters, speeding up feature generation by 5.08x (-80.3% latency). Drove INR 180 Cr (\$21M+) in annualized loan disbursal uplift, improved credit approval rates by 150 bps, and saved \$550,000+ in annualized cloud compute costs.

Technical Skills

- **Programming Languages:** Python, Scala, Kotlin, SQL, C/C++, Bash
- **Machine Learning & Deep Learning:** PyTorch, Hugging Face, Transformers, QLoRA, LightGBM, XGBoost, CatBoost, Scikit-learn, Ray, BiLSTM-CRF
- **Generative AI & LLM Systems:** Multi-Agent Underwriting, Few-Shot Prompting ($k = 2$), Dual-Pass Discriminator Review, Chain-of-Thought Verification, Deterministic Auto-Correction, Modular Prompt Registries, Adversarial Drift Auditing
- **Big Data & Distributed Systems:** Apache Spark, Databricks (100-node clusters), Apache Kafka, Delta Lake, Unity Catalog, Ray Distributed Tensors
- **MLOps & Cloud Infrastructure:** AWS (S3, EFS, EC2), MLflow, FastAPI, Reactive Kotlin, PostgreSQL, Docker, GitLab CI/CD, System Design, REST APIs
- **Optimization & Mathematics:** Submodular Information Optimization, Cobb-Douglas Elasticity Modeling, Compressed Sparse Row (CSR) Trie Data Structures, Statistical Hypothesis Testing

Professional Experience

Navi Technologies

Machine Learning Engineer 2

Bengaluru, India

July 2023 – Present

- **Algorithmic Template Selection & Submodular Optimization:** Solved a critical coverage bottleneck across 626M weekly SMS (where 53–69% are net-new templates weekly) previously throttled by rigid heuristic filters; formulated a novel Cobb-Douglas priority production function ($Priority = Importance^{1.2} \times [\ln(1 + SMS)]^{0.5} \times [\ln(1 + Users)]^{0.8}$) balancing 5-dimension local Gemma-4-12B rubric scoring with log-dampened population and user reach. Implemented a greedy submodular marginal-gain selection algorithm along the Pareto frontier, expanding weekly template selection capacity 20x (from 5k to 105k+ templates) and boosting SMS population coverage from 4.45% to 91.3% (+85.5% coverage gain) while capturing previously discarded high-intent financial alerts (loan overdues, disbursements, credit limits).
- **Automated LLM Annotation & Verification DAG (Terra & JEV):** Replaced manual human annotation queues (700 templates/day capacity, 20.1% baseline accuracy) with an asynchronous 10-task Databricks DAG orchestrating zero-shot hierarchical classification (gpt-5.6-terra), category-scoped few-shot entity extraction ($k = 2$), dual-pass fast discriminator review (deepseek-v4.1-flash), a zero-cost deterministic rule engine, and reasoned chain-of-thought verification on modified templates. Surged template accuracy from 20.1% to 87.7% (90.0% on core tags), achieved 99.0% precision and 98.5% recall on Tier 1 financial entities (*amount*, *balance*, *accountNo*), reduced tag errors by 74.3%, and eliminated review stall rates from 33.3% to 0% (100% automated completion).
- **High-Performance Compressed Sparse Row (CSR) Trie Engine:** Eliminated reflective $\mathcal{O}(T \times W)$ linear template scanning across thousands of classes by engineering an $\mathcal{O}(W)$ Flat-Array Compressed Sparse Row (CSR) Trie in Scala using contiguous primitive arrays (`childOffset[]`, `edgeLabel[]`, `edgeTarget[]`, `sChild[]`, `nodeTemplate[]`) loaded via bulk binary memory mapping. Scaled across 8.39 million templates and 138,721 buckets (237.5M nodes) on a 100-node Databricks Spark cluster (`r-fleet.2xlarge` workers), cutting feature generation latency by 80.3% (5.08x speedup from 1,586ms to 312ms), ledger extraction latency by 77.8% (4.51x speedup), and gateway wall-clock latency by 89.6% (9.61x speedup) with 100% exact parity across 249,689 test samples.
- **LLM-Based Underwriting Journey & Conversational Guidance:** Developed an end-to-end guided re-verification platform for rejected loan applicants, coupling an in-house ITR & GST parser (IGNIS) with a multi-agent LLM underwriting engine and multilingual chatbot for personalized document retrieval. Orchestrated a dynamic 3–5 model ensemble generating 150+ features, delivering a +36 bps credit approval rate uplift, INR 180 Cr (\$21M+) in annualized loan disbursal uplift, and INR 10–15 Lakhs in monthly operational cost savings.
- **Credit Risk & Income Predictive Modeling:** Advanced Navi's flagship Prism credit risk scorecard by engineering novel

temporal features and replacing univariate retraining pipelines with a multi-modal gradient boosting architecture, directly improving credit approval rates by 150 bps. Built an automated income prediction model using XGBoost and LightGBM, increasing prediction accuracy within a 10% tolerance threshold by 6 percentage points.

- **Distributed MLOps, Real-Time Serving & Cloud Cost Optimization:** Co-architected “SPADE Real Time,” a 3-tier reactive Kotlin microservice suite serving SMS, app, and location features at sub-20ms P99 latency at 10,000+ RPS over Kafka and PostgreSQL. Built an automated zero-drift feature store with 99.99% dev/prod consistency that ran 50% faster at half the infrastructure cost. Designed a Databricks compute regression monitor and access attribution framework that secured \$550,000+ in annualized AWS savings; optimized feature selection jobs (Boruta, RFE, IV) using Ray and distributed tensors to slash compute time and training costs by 80%.

Projects

Personal Project

Auto Code Sequence Generator

Oct 2022 – Apr 2023

- Built an Attention-based Transformer model to synthesize production-ready UI code from natural language prompts and wireframe layouts; designed a customized token vector mapper supporting seamless cross-compilation into React and Angular frameworks.

Hardware & Deep Learning Project

Visual Aid for Visually Impaired

Mar 2022 – Apr 2022

- Developed an edge embedded IoT system using ESP32 and Arduino-Nano transmitting wireless camera feeds to an inference server running YOLOv3 and Transformer models for real-time scene understanding and auditory field description trained on MS-COCO.

Deep Learning Project

AI Image Caption Bot

Mar 2021

- Developed a deep multimodal sequence-to-sequence captioning model combining CNN visual feature extractors with recurrent LSTM networks trained on the Flickr30k benchmark (1.5M parameters).

Education

Netaji Subhas University of Technology

New Delhi, India

Bachelor of Technology in Computer Science

Aug 2019 – May 2023

- **CGPA:** 8.72 / 10.0 (Graduated with Distinction).
- **Scholarship:** Recipient of the prestigious CVSPK Merit Scholarship (2020): 100% Tuition Fee Waiver for academic excellence.

Research Publications

Indian Journal of Computer Science

Early Detection of Covid-19 using Machine Learning

May 2021 – July 2021

- Utilized a fine-tuned ResNet-50 deep architecture to extract visual radiographic features from chest X-rays, achieving 99% validation accuracy in differentiating COVID-19 pathologies from normal and bacterial pneumonia scans. (DOI: [10.17010/ijcs/2022/v7/i1/168953](https://doi.org/10.17010/ijcs/2022/v7/i1/168953))

arXiv Preprint

Classification of Skin Cancer Images using Convolutional Neural Networks

Apr 2021 – May 2021

- Developed a two-stage deep learning framework leveraging XceptionNet for lesion image segmentation and a custom Dense CNN for binary benign vs. malignant classification, achieving an 86% validation accuracy. (arXiv: [2202.00678](https://arxiv.org/abs/2202.00678))