KARTIKEYA AGRAWAL

AI / ML Engineer | LLM systems, agents, evaluation and applied machine learning
Jaipur, India | +91 9352916929 | kartikay3@outlook.com
kartikeyaagr.com | linkedin.com/in/kartikeyaagrawal | github.com/kartikeyaagr

EDUCATION

Ashoka University Sonipat, India
B.Sc. (Hons) in Mathematics and Computer Science; Concentration in Entrepreneurial Leadership and Strategy May 2026
Relevant Coursework: Machine Learning, Reinforcement Learning, Computer Vision, Data Structures & Algorithms

EXPERIENCE

Trisula Capital Jaipur, India
Investments Intern May 2026 – Sep 2026
- Covered 20+ listed Indian small- and mid-caps for a long-only fund, from initial screening to investment-ready research.
- Met management teams and conducted plant visits to assess strategy, capital allocation and execution risk; built market-sizing models and financial-statement analyses to test each thesis.
- Built the fund's internal research and portfolio system, one place for the team to track its ~40-company coverage.

Aether AI Gurgaon, India
AI Associate; previously AI Engineering Intern Jul 2025 – Jan 2026
- Developed a multi-agent LLM framework with tool-calling over client ERP/CRM data, and an MCP server generating real-time KPI insights for downstream AI workflows.
- Engineered a FastAPI integration of a knowledge graph into the client dashboard, strengthening its core RAG application, and built a Google-API pipeline syncing approved sales orders into the client's ERP/CRM.
- Built a Streamlit demo of a voice-based AI assistant that directly helped secure a new client project, then built and deployed its React + Next.js web app.

Rays Power Infra Pvt. Ltd. Jaipur, India
Data Science Intern Jun 2025 – Jul 2025
- Engineered an automated LangChain pipeline for multimodal data in a RAG system, improving team efficiency by 25%.
- Automated goods-receipt notifications from BigQuery (SQL) into SAP S/4HANA ERP, reducing manual effort by 50%.
- Built a Google Workspace and Gemini chatbot for internal email and calendar, improving productivity by 20%.

PROJECTS

SmartRoute – LLM Inference Router Feb 2026 – Sep 2026
- Routed each request to the cheapest capable model using logistic-regression triage over frozen MiniLM embeddings, with thresholds derived from a cost break-even formula rather than hand-tuned.
- Built a live cost-and-accuracy harness and MMLU evaluation against always-cheap, always-middle and always-frontier baselines; on a 200-question corpus the cheap tier matched frontier accuracy at 11.8x lower cost.
- Shipped an OpenAI-compatible FastAPI server on LiteLLM across Together AI, Anthropic and Groq, with pgvector semantic caching, a Postgres decision log, Prometheus metrics, OpenTelemetry traces and 202 tests.

Folio – Fund Operating System May 2026 – Sep 2026
- Unified NSE filings, pricing, XBRL fundamentals, shareholding and insider/pledge disclosures since 2015 into one dataset, with a portfolio book and alerting on top.
- Put six task-specific AI agents behind a human review gate, with a loop feeding analyst edits into future outputs.
- Ran it on a single AWS EC2 server with private retrieval, keeping proprietary research on the fund's own infrastructure.

Quark – LLM from Scratch Apr 2026
- Built a decoder-only LLM in PyTorch with a LLaMA-3-style architecture (grouped-query attention, RoPE, SwiGLU) and a byte-level BPE tokenizer, with presets from 16M to 1B parameters.
- Added INT8/INT4 weight quantization (4.0x/7.9x smaller) and a compressed 3-bit KV cache using ~3.4x less memory.
- Benchmarked prefill throughput, decode latency and peak VRAM; shipped training and generation CLIs with 56 tests.

fast-laya-compaction – Claude Code Plugin Sep 2026
- Built and published a Claude Code plugin that compacts long sessions without summarizing, keeping every user and assistant message verbatim and trimming only stale tool calls.
- Paired a deterministic rule for superseded calls with a local 421M-parameter classifier that keeps, truncates or drops old tool calls in one batched pass.
- Cut context size by 27–33% on real sessions in 19–47 s versus 79 s for built-in compaction, falling back safely on failure.

SKILLS

Languages: Python, TypeScript, SQL
LLMs and Agents: RAG, multi-agent systems, tool-calling, MCP, prompt engineering, LLM evaluation, LangChain
Machine Learning: PyTorch, Hugging Face Transformers, scikit-learn, fine-tuning, quantization, computer vision, NLP
Backend and Data: FastAPI, REST APIs, PostgreSQL, pgvector, BigQuery, Pandas, data pipelines, React, Next.js, Node.js
Infrastructure: Docker, AWS EC2, Google Cloud, Prometheus, OpenTelemetry, MLflow, Weights & Biases, pytest, Git
Finance: Equity research, financial statement analysis, valuation, market sizing, algorithmic trading, backtesting