Aayush Kumar

Chicago, USA
+1 (312) 375-5603 | akuma102@uic.edu | https://linkedin.com/in/aayushakumars/
https://github.com/aayushakumar/

Summary

Senior-level ML Engineer with deep expertise in building productiongrade AI security and retrievalaugmented generation pipelines. Delivered an agentic security copilot that kept false positives under 3% while automerging over 30% of remediation PRs and processing 50+ PRs weekly within a 24hour SLA. Previously led ZDX Copilot, cutting hallucinations by 18% and boosting incident detection by 30%, while slashing memory usage from 220 GB to 52 GB and reducing compute costs sixfold. Seeking to apply this blend of ML, performance optimization, and security automation to accelerate robust, scalable AI solutions for the organization.

SKILLS

- Python, C/C++, Java, Perl, PyTorch, TensorFlow, JAX (familiar), Transformers, Retrieval-Augmented Generation (RAG), LangChain, FAISS, Gradio, Kubernetes, Docker, Spark, Distributed Systems, REST, gRPC, Performance analysis and optimization, Profiling (memory/runtime), Data optimization, SQL, Postgres, Git, OPA, Rego, CloudTrail, Steampipe, Policy-based access/remediation, Prompt engineering, LLM guardrails, Graph-based recommendation (Louvain), Containerization, Monitoring/metrics/alerts, AWS (GitHub/AWS integrations), Security-focused ML systems

WORK EXPERIENCE

HyperSentry (AI Security Copilot) | ***ML Systems Engineer (Co-Founder)*** **May 2025 - Present**

- Built a security copilot using OPA/Rego policy checks and retrievalaugmented generation over security knowledge bases, automatically creating remediation pull requests in GitHub and AWS, which streamlined the remediation process and reduced manual effort
- Held false positives under 3% while auto-merging 30%+ of fixes; processed 50+ PRs/week with 24h remediation SLA.
- Optimized retrieval + reasoning pipeline (embedding cache, batched inference, parallel I/O) to keep E2E <800 ms under typical load.
- Integrated Slack and CLI interfaces for naturallanguage queries and autogenerated CloudTrail and Steampipe queries to clarify blast radius, enabling faster incident analysis

Z*scaler** | ***Machine Learning Engineer **Feb 2022 - Aug 2024**

- Led ZDX Copilot (RAG + LLM Guardrails): tuned retrieval and prompt-routing to cut hallucinations by 18% and improve ISP incident detection by 30% in production.
- Profiled ML & data services (memory + runtime) using Python/pandas profilers and app-level tracing to reduce peak memory from 220 GB to 52 GB and 6× compute cost reduction.
- Designed a hybrid Spark + Kubernetes architecture to 4× training data capacity for large tenants, enabling distributed processing without degrading SLOs.
- Built a graph-based policy recommender (Louvain clustering on sequential access logs) adopted for 80% of the top 100 policy recs.
- Containerized and instrumented ML services (K8s, metrics/alerts) for reliable rollout across tenants.

PROJECTS

[PATENT] LLM-accelerated security classification | ***US-20250119432-A1***

- Developed a patented LLM-based content classification system that processes and categorizes 10,000+ documents weekly for compliance, improving compliance and data governance.
- Boosted threat detection accuracy by 25% through optimization of the LLM-based classification system. Optimized tokenization (Tiktoken) to cut processing time by 30%.

TariffWhisperer: AI-Powered HTS Classification Assistant

- Built an end-to-end RAG pipeline over 10,000+ CBP rulings with FAISS for sub-500 ms semantic retrieval.
- Doc triage engine with FAISS (dense) + BM25 hybrid, sub-500 ms retrieval, justification summaries, and guardrails; framed as relevance ranking + explainable snippets for reviewers.
- Conducted prompt engineering to reduce hallucinations and improve justification relevance. Performed model- and graph-level optimizations (op fusion, tokenization optimizations) to reduce inference latency.

Phishing Adversary Framework: Attacker–Defender Loop

- Trained a BERT-based detector against an RL-based adversary that generates minimally evasive phishing emails.
- Achieved a 25 percentage-point absolute gain in detection under adversarial attack versus a static baseline, while preserving 96.8% accuracy on clean data.

EDUCATION

University Of Illinois Chicago (UIC) | ***Master of Science, Computer Science*** **Aug 2024 - May 2026**

Manipal University | ***Bachelor of Technology, Information Technology*** **Aug 2018 - Aug 2022**