

Aman Singh Thakur

Machine Learning Engineer - Agentic AI, RAG & Foundational Research

5+ Years · Seattle, WA, USA

✉ singh96aman@gmail.com | 🌐 singh96aman.me | 💻 in/singh96aman | 🌐 singh96aman | 📄 Scholar



Professional Experience

Machine Learning Engineer 2, Amazon - Seattle, USA

Jul 2024 – Present

Agentic AI · RAG · Foundation Model Evaluation · Model Customization · Reinforcement Learning

- **Opinionated RAG (Seller Assistant):** Integrating opinion-aware retrieval into Amazon's Seller Assistant, serving ~10M selling partners with responses grounded in Seller Central and Reddit while preserving representative sentiment.
- **Foundation Model Evaluation & Multi-Turn RL:** Built SageMaker's model-evaluation service enabling base-vs-custom comparison with LLM-as-judge, reward, pass@k, and trajectory metrics; extended to serverless multi-turn RL for tool-use and reasoning tasks.
- **Agentic Model Customization:** Delivered an Agents PySDK/Skill workflow automating fine-tuning, evaluation, and deployment on SageMaker; demoed at re:Invent 2025, scaled to 50+ customers, integrated via A2A/MCP across 10+ base models (SFT, DPO, RLVR).
- **Fraud Model Revamp (Mechanical Turk):** Overhauled the rule engine and retrained a PyTorch network to counter data drift, improving fraud-detection accuracy from 82% to 97% across global marketplaces and defending against \$25M+ in potential losses.

Software Engineer 2, Goldman Sachs - Bengaluru, India

Oct 2020 – Jul 2022

Applied ML · Statistical Modeling · Anomaly Detection · Python, SQL · Data Engineering

- **Credit Risk Modeling:** Built credit-scoring models using PCA, RFE, and Spectral Embedding feature selection, achieving 95%+ accuracy across 10K+ loan applicants.
- **Risk Exposure & Anomaly Detection:** Deployed real-time anomaly detection monitoring \$500Bn+ firm exposure across 100+ clients, flagging trading irregularities for risk mitigation.
- **Regulatory Data Infrastructure:** Engineered multi-language pipelines (Java, Python, C++) validating 100+ data-quality contracts and processing 20Bn+ daily records across data-lake, Sybase, and DB2.

Software Engineer, Morgan Stanley - Mumbai, India

Jul 2019 – Sep 2020

NLP · Java · Spring · SQL · Data Pipelines

- **NLP Trade-Ingestion Platform:** Built an NLP pipeline extracting trade signals (tickers, buy/sell, quantities) from CSV, XML, and email sources, onboarding 50+ institutional clients.
- **Autonomous Trade Processing:** Reduced trade-processing time 40% with an Angular UI over a Spring backend for customizable processing templates.

Publications & Research

Community - Grounded Memory: Cold-Start Personalization for LLM

Sep 2026

Assistants from User Attributes

Amazon · Personalization, Cold-Start, RAG · Under submission, EACL Industry Track 2027

Personalizes cold-start users from a community-grounded prior that predicts their opinions from first-contact attributes, injected as a short natural-language memory; on a production e-commerce assistant, a blinded LLM-judge panel preferred its first-turn replies over a no-memory baseline on 89% of questions.

Training Leaves Traces: Centered Residual Signatures for Language Model

Aug 2026

Lineage Verification

Amazon · Model Lineage, LLM Provenance · Under submission, ACL ARR Aug 2027 · [arXiv:2608.14929](https://arxiv.org/abs/2608.14929)

Residual fingerprint achieving AUROC 1.0 distinguishing fine-tuned, LoRA-merged, and quantized descendants from independent models; 76× faster than baselines; validated across 6 LLM families.

- WARP: Wasserstein - Aligned RAG for Population Opinions**

Amazon · RAG, Information Retrieval · EMNLP'26 Industry Track (Oral) · [arXiv:2608.22859](#)

Aligns retrieved evidence with population opinion distributions, reducing distributional error by 45–81% across three domains within sub-second latency.

Jun 2026
- Retrieval - Augmented Generation Must Move Beyond Factual Grounding to Represent Diverse Opinions**

Amazon · RAG, Information Retrieval · INLG'26 Main Conference (Oral) · [arXiv:2604.12138](#)

Proposed O-RAG to preserve corpus-level opinion distributions while maintaining semantic similarity; surveyed 34 RAG benchmarks (none preserve population opinion); human study showed 79% preference for opinion-enriched content.

Apr 2026
- Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges**

UMass Amherst - Meta · LLM Evaluation, Generative AI · [ACM GEM \(250+ citations\)](#)

Benchmarked 13 LLM judges on 9 models; exposed systematic biases (leniency, length sensitivity) with scores differing up to 5 points from human ratings.

Jun 2024
- Baler: Deep Autoencoders for Scientific Data Compression**

Google Summer of Code 2023 · CERN-HSF / Lund University · [ResearchGate](#)

Extended Baler with 2D/3D convolutional autoencoders achieving 95%+ reconstruction at 200%+ compression; added error-bounded regularization.

Oct 2023

Education

- University of Massachusetts Amherst** - Amherst, USA

M.S. in Computer Science

Sep 2022 – May 2024

CGPA 3.9/4.0
- Manipal Institute of Technology** - Manipal, India

B.Tech in Computer Science & Engineering

Jun 2015 – Jun 2019

CGPA 8.4/10

Academic Service, Mentorship & Peer Review

- **Sep 2026** – Reviewing 2 papers for the ACL ARR August 2026 cycle.
- **Aug 2026** – Reviewing 4 papers, including as an Ethics reviewer, for INLG 2026.
- **Jul 2026** – Reviewed 5 papers plus 2 as an Ethics reviewer for the ACL ARR May 2026 cycle.
- **Jun 2026** – Reviewed 7 papers for the Amazon ML Conference (AMLC 2026) and 7 proposals for Amazon Analyticon 2026; program-committee reviewer for EMNLP 2026 Industry Track.
- **Ongoing** – 1:1 Mentorship on Topmate platform with 50+ engineers to help them succeed in their machine learning careers.

Awards & Honors

- 3rd Place, Amazon Worldwide Stores GenAI Fair** - for the Opinion-Aware RAG (O-RAG) poster.

Sep 2026
- Two-time Google Summer of Code Grantee** - \$9,000 in grants from Google to build open-source CERN-HSF libraries pyAWAKE and Baler.

2019, 2023

Technical Skills & Certifications

ML & GenAI	PyTorch, RAG, Information Retrieval, LLM Evaluation, LLM-as-Judge, vLLM
Foundational Model	SFT, DPO, RLVR, LoRA/PEFT, Reinforcement Learning, Reward Modeling, Multi-Turn Evaluation
Agentic Systems	Agents SDK, MCP, A2A/ACP, LangChain, Agentic Workflows, Tool-Use
Software Engineering	Python, Java, C/C++, Go, Rust, SQL, Spark, Flink, Kafka, AWS, SageMaker, Bedrock, Docker, Kubernetes
Certifications	AWS Certified Machine Learning Engineer - Associate