

Aman Singh Thakur

Machine Learning Engineer - Agentic AI, RAG, and Foundational Research | 5+ Years

Seattle, WA, USA | +1-4134726405 | singh96aman.me | singh96aman@gmail.com | [linkedin.com/in/singh96aman](https://www.linkedin.com/in/singh96aman) | [Google Scholar](https://scholar.google.com/citations?user=...)

PROFESSIONAL EXPERIENCE

Amazon, Seattle, USA

July 2024 - Present

Machine Learning Engineer 2 | Agentic AI, RAG, Foundation Model Evaluation, Model Customization, Reinforcement Learning | [TechCrunch](#)

- *Opinionated RAG (Amazon 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.
- *Agent-Driven Model Customization*: Delivered 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, supporting **10+ base models** across Supervised Fine-Tuning, Direct Preference Optimization, and Reinforcement Learning with Verifiable Reward.
- *Fraud Model Revamp (AWS Mechanical Turk)*: Overhauled the rule engine and retrained a PyTorch neural network to address data drift, improving fraud detection accuracy from **82% to 97%** across global marketplaces and defending against **\$25M+** in potential losses.

Goldman Sachs, Bengaluru, India

Oct 2020 - Jul 2022

Applied Scientist 2 | 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** for regulatory reporting across data-lake, Sybase, and DB2.

Morgan Stanley, Mumbai, India

Jul 2019 - Sep 2020

Software Engineer | NLP, Java, Spring, SQL, Data Pipelines

- *NLP Trade-Ingestion Platform*: Built NLP pipeline extracting trade signals (tickers, buy/sell, quantities) from CSV, XML, and email sources, onboarding **50+ institutional clients**.

PUBLICATIONS & RESEARCH

Training Leaves Traces: Centered Residual Signatures for Language Model Lineage Verification

Aug 2026

Amazon | Model Lineage, LLM Provenance, Weight-Space Analysis | [Preprint \(arXiv:2608.14929\)](#)

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

WARP: Wasserstein-Aligned RAG for Population Opinions

Jun 2026

Amazon | RAG, NLP, Information Retrieval | [EMNLP'26 Industry Track \(Oral Accept\)](#) | [arXiv:2608.22859](#)

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

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

Apr 2026

Amazon | RAG, NLP, Information Retrieval | [INLG Main Conference'26 \(Oral Accept\)](#) | [arXiv:2604.12138](#)

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

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

Jun 2024

UMass Amherst–Meta Collaboration, Remote, USA | NLP, LLM Evaluation, Generative AI | [ACM GEM² \(250+ citations\)](#)

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

Baler - Deep Autoencoders for Scientific Data Compression

Oct 2023

Google Summer of Code 2023 · CERN-HSF / Lund University | PyTorch, Autoencoders, Spark | [ResearchGate](#)

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

EDUCATION

University of Massachusetts Amherst, USA

Master of Science in Computer Science

Sep 2022 - May 2024

CGPA: 3.9/4.0

Manipal Institute of Technology, India

Bachelor of Technology in Computer Science Engineering

Jun 2015 - Jun 2019

CGPA: 8.4/10

TECHNICAL SKILLS

- | | |
|---|---|
| <ul style="list-style-type: none">• ML & GenAI• Model Customization• Cloud & Agent Systems• Data & Languages | <p>PyTorch, RAG, Information Retrieval, LLM Evaluation, LLM-as-Judge, Reinforcement Learning</p> <p>SFT, DPO, RLVR, LoRA/PEFT, Agentic Workflows, Multi-Turn Evaluation</p> <p>AWS, SageMaker AI, Bedrock, Agents SDK, MCP, A2A, LangChain, Docker, Kubernetes</p> <p>Python, Java, C/C++, SQL, Go, Spark, Flink, Kafka, Redshift, S3</p> |
|---|---|