# Rishabh Sood

LinkedIn: linkedin.com/in/rishabhsood2
GitHub: github.com/RishabhSood
Email: sood.rishabh@hotmail.com
Mobile: +91-8920492376
Blog: rishabhsood.github.io

## Education

- **Georgia Institute of Technology** — Atlanta, GA (Remote)
  *M.S. Computer Science, Machine Learning Specialization (GPA: 4.0/4.0)* — *2025 - Present*
- **Thapar Institute of Engineering & Technology** — Patiala, India
  *B.E. Computer Science & Engineering (GPA: 9.92/10.0)* — *July 2019 - June 2023*

## Experience

- **Apple** — Hyderabad, India
  *Machine Learning Engineer* — *Jan 2024 - Present*
  - **Support App Chat (Generative Experience)**: Engineered one of Apple's first generative multi-turn conversational experiences, providing an instant 24/7 support channel for **millions** of users, with a modular architecture supporting tool invocation (warranty checks, scheduling) and strict response validation.
    - Architected a **multi-agent orchestration layer** with clearly defined interfaces, fallback strategies, and latency-aware routing; reduced **p95 latency** by **75%** by replacing rule-based agents with a **fine-tuned Qwen-14B** model.
    - Engineered a production-grade **RAG pipeline** grounded in proprietary knowledge sources, emphasizing correctness guarantees, hallucination mitigation, and schema-constrained outputs under high request throughput.
    - Implemented safety and policy enforcement layers with runtime checks and fallback handling to ensure deterministic behavior and compliance across all generated responses.
  - **Apple Developer Search**: Led geographical expansion to multiple regions/languages, optimizing storage costs by **90%** (10x reduction) via Solr/OpenSearch optimizations without compromising search quality.
  - **Knowledge Hub**: Architected a centralized ETL pipeline (MQs/APIs → Cleaning & Embedding → Vector/Graph DBs) with a custom plugin architecture, enabling dynamic pipeline composition and faster onboarding of new data sources.
  - **Tech Stack:** Python, LLM Fine-tuning (Qwen), RAG, GraphRAG, LangChain, Textgrad, MCP, Knowledge Graphs, FastAPI, Solr, OpenSearch, Neo4J, Docker.
- **Goldman Sachs** — Hyderabad, India
  *Analyst* — *Jul 2023 - Dec 2023*
  - **Infrastructure Migration**: Designed and implemented core components of a **Kubernetes**-based migration for the central App Store, enabling automated UAT and improved deployment reliability.
  - **Data Aggregator**: Architected a highly configurable system to automate Kubernetes job scheduling for data lake onboarding, significantly reducing manual overhead and infrastructure costs.
  - **Tech Stack:** Java, Spring Boot, Kubernetes, Docker, GitLab CI/CD, Gherkin, Cypress.
- **Apple** — Hyderabad, India
  *Software Development Engineer (Intern)* — *Jan 2023 - Jun 2023*
  - **Intelligent Recommendation Service**: Developed a service using the dynamic Apple Support Knowledge Base (real-time sync), reducing response time from minutes to sub-second latency.
  - Integrated generative responses and sentiment analysis to improve user support outcomes.
  - Innovated a centralized Python package to reduce code redundancy across internal teams.
  - **Tech Stack:** Python, PyPI Packaging, FastAPI, MongoDB, OpenSearch, Solr, Amazon EKS.
- **Goldman Sachs** — Bengaluru, India
  *Summer Analyst (Intern)* — *Jun 2022 - Jul 2022*
  - **Emergency Notification Service**: Developed and deployed a global notification service, replacing a legacy 3rd party vendor solution to save **USD 35-40K/annum**.
  - Designed the system for extensibility, allowing new regions to be onboarded via simple configuration changes.
  - **Tech Stack:** Golang, Kubernetes, Prometheus, Grafana, BigQuery, Docker.

## Technical Skills

**Languages**: Python, Rust, C/C++, Java, Golang, SQL
**ML & Frameworks**: PyTorch, DeepSpeed, Pandas, NumPy, Scikit-learn, LangChain, Textgrad
**Backend & Systems**: FastAPI, Spring Boot, gRPC, Rust (RIG, Poem), Docker, Kubernetes (EKS)
**Databases & Search**: PostgreSQL, MongoDB, Redis, OpenSearch, Apache Solr, Qdrant, Neo4J

## Trainings & Certifications

- **Generative AI with LLMs** (Coursera), **Deep Learning Specialization** (DeepLearning.AI), TensorFlow in Practice, AWS Fundamentals, Architecting on AWS, Amazon EKS, Google Code-In