NIRANJANA AMBADI

AI Engineer, Researcher & Founder

Portland, Oregon • +1-617-487-9895 • niranjana13@gmail.com • LinkedIn • GitHub • U.S. Permanent Resident

SUMMARY

AI Engineer, Researcher & Founder with a PhD in information theory and deep hands-on experience building production RAG systems, agentic retrieval pipelines, and evaluation frameworks. Proven ability to take state-of-the-art research techniques from ideation to shipped product — across knowledge base construction, query understanding, retrieval ranking, and automated metrics. Experienced defining task-specific evaluation datasets, driving rigorous offline/online experimentation, and leveraging AI coding agents to accelerate research iteration. Strong ownership mindset with a bias for action and cross-functional collaboration.

CORE COMPETENCIES

Agentic Retrieval & RAG: Agentic retrieval pipelines, query rewriting, hybrid search (BM25 + dense), re-ranking, RAG, vector databases (FAISS, S3 Vectors), semantic search, knowledge base construction

Evaluation & Metrics: Task-specific evaluation dataset construction, offline/online experimentation, nDCG, precision/recall, regression detection, automated benchmarking, A/B testing, production monitoring

ML Research → Production: State-of-the-art research integration, LLM prompt development, agent scaffolding, PEFT/LoRA fine-tuning, DPO/RLHF, transformer architectures, HuggingFace

ML Pipelines & Infrastructure: PyTorch, AWS (Bedrock, AgentCore, Lambda, S3, EC2), FastAPI, multi-GPU distributed training, end-to-end pipeline engineering

Languages & Tools: Python (primary), Java, SQL, JavaScript, C++, R; AI coding agents (Claude, Copilot), LangChain

WORK EXPERIENCE

Founder & CEO | LawMate (Private Limited) *Jan 2026 – Present | Portland, OR / Kerala, India*

AI-Powered Legal Workspace Custom-Designed for the Kerala Bar — Agentic Retrieval, RAG, Evaluation Infrastructure

- Built a production agentic retrieval system over a 35K+ document legal knowledge base on AWS S3 Vectors — incorporating query understanding, semantic search, hybrid retrieval, and grounded generation with strict citation accuracy requirements.
- Designed and implemented a three-tier evaluation pipeline: lightweight model for query triage and intent classification, grounded generation model for response synthesis, and deterministic post-processing with anti-hallucination guardrails — with measurable regression detection across model updates.
- Engineered multi-agent orchestration (Legal Chatbot, Drafting AI, Case Prep Agent) with structured tool use, stateful multi-turn reasoning, and modular routing — built from scratch on AWS, shipped to production from architecture to launch.
- Developed 64 document-type LLM prompts with statute crosswalk tools (BNS/BNSS/BSA) and cause list parsing — iterating prompts using AI coding agents to accelerate experimentation and improve retrieval relevance across complex legal queries.
- Constructed curated evaluation datasets for legal query types; defined task-specific metrics for retrieval quality, citation accuracy, and drafting correctness — driving continuous improvement across the knowledge retrieval stack.

Associate (SDE) | Goldman Sachs *Jul 2022 – Oct 2024 | Salt Lake City, UT*

Private Wealth Management — High-Performance Systems, Distributed Infrastructure

- Built automated compliance audit pipeline for 50,000+ accounts using intelligent batching and parallel processing in Java — 60× throughput improvement with full observability and stakeholder reporting dashboards.
- Developed production APIs with Geode Cache (distributed in-memory data grid); improved data access latency by 40%. Achieved 99.5% platform uptime and reduced MTTR by 30% through proactive monitoring and Tier 2/3 support.

Graduate Research Assistant | Boston University *Jan 2021 – May 2022 | Boston, MA*

ML Research — Novel Classifiers, Robust Statistics, Heavy-Tailed Distributions

- Developed Stable-QDA: novel robust classifier using α-stable distributions; designed comprehensive evaluation framework with benchmarks across synthetic and real-world datasets (HTRU2, Credit Card Fraud, NetML) — presented at ML Week Europe, Munich, Nov 2024.

Applications Engineer | Oracle *Jun 2013 – Jul 2014 | Bengaluru, India*

Fusion CRM — Enterprise Java, Oracle ADF, Performance Testing

- Engineered scalable components for AppComposer (Fusion CRM) in Java/Oracle ADF; reduced production defects by 35% through rigorous testing pipelines.

SELECTED PROJECTS

Generative Retrieval with Preference-Optimized Re-ranking | PyTorch, HuggingFace, AWS

- Built three-stage IR pipeline (BM25 → Dense Retrieval → LLM Re-ranker) with DPO-trained Mistral-7B using PEFT/LoRA; improved nDCG@10 by 22% on BEIR benchmarks.
- Defined task-specific evaluation datasets and metrics; implemented distributed multi-GPU training on AWS with end-to-end offline evaluation framework tracking regression across model iterations.

Agentic Job Search System | Python, Streamlit, Qualcomm AI Playground, SERP API

- Built an end-to-end agentic pipeline leveraging LLM inference APIs to rank and score job postings against user-defined requirements — demonstrating long-running agent scaffolding, data pipeline construction, and cross-platform LLM deployment.
- Integrated real-time data ingestion (SERP API), SQLite persistence, and automated evaluation/alert delivery — full pipeline from data collection through automated metrics to delivery.

EDUCATION

PhD, Electrical Communication Engineering | Indian Institute of Science, Bengaluru *2014 – 2020*

Dissertation: Selected Problems in Network Coding Using Tools From Linear Algebra and Matroid Theory

MS, Computer Information Systems | Boston University *2021 – 2022*

Certificate, Data Science & ML | MIT Schwarzman College of Computing *2022*

PyTorch for Deep Learning Professional Certificate | DeepLearning.AI *2026*

PUBLICATIONS & PRESENTATIONS

- **Stable-QDA:** "Likelihood over Estimation: Robust Quadratic Discriminant Analysis for Heavy-Tailed Distributions with Theory and Evidence" — Accepted, ICML 2026; presented at ML Week Europe, Munich, Nov 2024. Built fully scikit-learn compatible robust classification library; designed and ran rigorous offline experiments to validate theoretical consistency and insensitivity propositions — demonstrating research-to-production execution.
- **Silver Reviewer Award, ICML 2026**; Reviewer, NeurIPS 2026.
- **Network Coding Research** — 7 peer-reviewed publications (IEEE ISIT, IEEE WCNC, IEEE conferences); reviewer for IEEE ISIT, IEEE WCNC.