

# GIRISH NATHAN

Chennai 600090 | +91 72000 96870 | girish.nathan@gmail.com

## PROFESSIONAL SUMMARY

Senior AI/ML Technology Leader (15+ years) with deep experience in building and leading high performing AI ML teams. Deep expertise in statistical physics, ML at scale, ML platforms, and delivering measurable business impact on a timely schedule. Experienced speaker and lecturer.

## CORE COMPETENCIES

Leadership & Strategy: ML platform architecture, AI CoE development, MLOps pipelines, feature engineering at scale, A/B testing, LLMOps and LLM frameworks

Techniques: Reinforcement learning, deep learning, contextual bandits, graph neural networks, NLP, hierarchical topic models, privacy-preserving ML. Fluent in deep learning frameworks like Tensorflow, Pytorch and modern LLM end to end toolkits like LangChain, LLamaIndex, MCP and A2A

Stack: C, TypeScript, Python, R, Spark, Ray, FLAML, prompt engineering, agentic AI workflows, awk, sed, perl

Domains: Fintech, e-commerce, travel, legal tech, healthcare

## PROFESSIONAL EXPERIENCE

### AI & ML Technology Consultant — Self-Employed, Chennai, India (Aug 2024 – Present)

- End-to-end ML pipeline & experimentation (+15% CTR, +8% conversions) that was used by a clothing chain to improve rack stacking and improving conversions ; early experiments using an LLM-driven retrieval-augmented product recommender indicate a CVR lift of 23% over baseline. Product-product similarities were modelled using dense embeddings in pgvector (a vector DB similar to Pinecone / Weaviate / Milvus) and provided a really fast retrieval approach for an LLM-based ranker to improve upon
- Retirement portfolio optimizer (+4 bps vs NSE, 12 months) for a family firm based out of Chennai. Leveraged **retrieval-based analytics** and **smaller transformer models (FinBERT, DistilBERT)** to extract and normalize unstructured financial data streams
- Restaurant Multi-Armed Bandit framework being tested out to enable dynamic pricing and choice for customers while also optimising for riders, **integrated with a retrieval layer** for item embeddings and **fine-tuned smaller LLMs (LLaMA-2-7B, Mistral)** to enrich customer/rider feedback signals. We saw from this project that using even smaller models (Microsoft Phi\* series) did not hurt production metrics by much.
- Apollo Hospitals agentic matching; privacy-preserving updates; 2-site pilot. Designed a **RAG pipeline** over semi-structured EHR data using **embedding-based retrieval (FAISS) + LLM orchestration (GPT-4, fine-tuned BioBERT)**, with **agentic flows** powering real-time updates.

- End to end legal document processing for a family law firm in Chennai - this involved processing a 25000+ document corpus of legal contracts, payments schedules, contractor agreements(each several hundred pages long ) using a **combination of SOTA OCR techniques (dots, olmocr, Qwen's vision model) and SOTA PDF processing tools (pdfminer and pdfplumber)**. Improved document processing and summarisation time from 10s of documents a week to over 250 documents a week. Enabled higher accuracies than human analysts by embedding a human expert in the loop for the first two weeks of this work. **Used MCP servers with tool calling** for - document processing, topic modelling, change management in the same document(s) over time.

#### **Chief Data Officer — Tonik Bank, Chennai, India (Oct 2023 – May 2024)**

- Led 10 ML + 8 DE and delivered a series of lectures on basic AI and stats company-wide (75 engineers, 30 product and business)
- Loan recommender in prod (Feb 2024) to improve loan production and quality by 11% in 4 months
- Built Ray/FLAML platform from scratch for speeding up modeling (~10× faster model development from prototype to production)

#### **Senior Director of ML & AI — Razorpay, Bangalore, India (Sep 2022 – Sep 2023)**

- Built AI/ML CoE (team 14), delivered 3 talks in industry leading conferences and ran workshops for LLMs across engineering and data science orgs
- Payment routing using contextual bandits (+17% over baseline, prod) to enable effective, on-time payments rolled across all 5000 payment “points”
- RTO insurance models (+7% vs XGBoost, prod) to combat the issue of fraudsters gaming e-commerce platforms ; production models reduced fraud by 18% and impacted customer NPS positively by 22 points over a 4 month pilot
- LLM apps for internal chatbots for internal knowledge recovery

#### **Director of ML & Applied Research — Expedia Group, Seattle, WA (Aug 2020 – Jul 2022)**

- Led team of 28 across SEO/mobile/CRM/incentives and grew team from a size of 5 to 28 in over 2 years
- NLP/DL to improve content quality of our landing pages with a view to optimise for PageRank and a secondary metric around improving two-step CTRs(increased by 19% over incumbent) and CVRs (increased by 14.5% over incumbent) (in prod July 2020)
- Bandits with knapsacks to improve carousel performance by serving relevant heterogeneous recommendations (flight + activity + coupon) to customers. Went from prototype to

production in under 8 weeks and resulted in a \$88MM bottomline improvement just across the top 150 US cities in 3 months after launch (March 2022)

- GNNs for travel KG recommendations

#### **Principal Researcher — Expedia Group, Seattle, WA (Oct 2019 – Aug 2020)**

- Review mining (+9% dwell, +5% transactions) to improve review relevance for customers based on implicit and explicit inferences on customer types
- Contextual bandits to improve quality of hotel recommendations above the fold on the US pages (+8% increase in CVRs)
- Hotel reco via matrix approx + bipartite matching (+8% engagement)

#### **Senior Applied Scientist — Amazon, Seattle, WA (Sep 2018 – Sep 2019)**

- RL engine for brand ad effectiveness (500+ brands, 1000+ expansions)
- Single-touch attribution; improved brand discovery, and using detrimental point processes to balance relevance with discovery

#### **Senior Data Scientist — Microsoft, Bellevue, WA (Aug 2014 – Jul 2018)**

- Multi-armed bandit framework in production for improving Microsoft Band nudges (+12% satisfaction, prod May 2016)
- Hierarchical topic modeling to identify email topics and sub-topics and enable auto-folding → rolled out to 100MM+ users in May 2017
- Privacy preserving machine learning using fully homomorphic encryption
- 4 projects shipped (led 2 scientists + 4 engineers)

#### **Machine Learning Scientist — Amazon, Bangalore, India (May 2012 – Aug 2014)**

- Automated browse classification (\$80MM+ savings over 2 yrs, prod Aug 2013) where we helped build a multi-class, multi-label classifier for over 25000 browse nodes
- Used LSH similarity to identify substitute products at scale (latency < 500 ms) and reduced 3P seller price-gouging attempts by 40%

#### **Senior Research Engineer — Yahoo!, Bangalore, India (Sep 2011 – May 2012)**

- Bayesian bandits for ads (+30% CTR, India) ; contextual advertising in the Indian marketplace

#### **Research Geophysicist — ION Geophysical, Houston & Denver, USA (Jul 2006 – Jul 2011)**

- Inverse optimization for seismic/exploration
- Wavelet signatures via lasso/basis pursuit

## **EDUCATION**

Ph.D. in Physics — University of Houston (2006) ; focus : spatiotemporal pattern formations in nonequilibrium systems

## **PUBLICATIONS & PATENTS**

Several, available from both Google Scholar and on request. Patents : 2 USPTO granted, 7 applied for