NITIN MOHAN

mohannitin321@gmail.com LinkedIn +44 7436911217 Portfolio

Education

MS Artificial Intelligence, *University of Southampton* (UK) **Sep 2021 - Sep 2022**

BTech Computer Science Engineering, *GGSIPU* (India) **Aug 2014 - Aug 2018**

Technical Proficiency

PyTorch	LangGraph / LangSmith	RAG & Vector Retrieval	AWS (SageMaker, ECS, Cloudwatch)	Model Quantization & Optimization
Transformers	Agent Orchestration	LoRA / QLoRA	NLP / scikit-learn	Vector Databases

Professional Experiences

Product Data Scientist, The Stepstone Group **London, UK | Jun 2023 – Present**

- Architected and productionised a multilingual semantic search platform for a two-sided job marketplace, enabling candidate profile–job listing matching across global markets.
  - Developed transformer-based retrieval models and training pipelines using SageMaker and MLflow, deployed on AWS with CI/CD (Infrastructure-as-Code) and production monitoring via Datadog
  - Achieved sub-100ms P99 latency at millions of daily queries through INT8 quantisation and embedding caching
  - Maintained nDCG@10 within 1% of full-precision baselines while driving a projected +65% YoY revenue uplift on the core search entry point

- Developed scalable fine-tuning infrastructure for transformer embeddings across multilingual markets.
  - Trained SentenceTransformers with LoRA/QLoRA using SageMaker pipelines, incorporating hard-negative mining and evaluation harnesses to control semantic drift
  - Designed synthetic data curation pipeline using GLiNER agreement scoring & human validation gates
  - Achieved near-human model performance (within ~3pp) at ~10x lower annotation cost

- Engineered a production multi-agent onboarding platform spanning signup, job discovery, application, and re-engagement at scale.
  - Designed agent architecture and inter-agent communication across LangGraph-based, stateful workflows
  - Implemented behavioural drift monitoring and Responsible AI guardrails aligned to EU AI Act requirements to ensure robustness in production
  - Improved application completion rates by +26%

- Applied regression discontinuity to evaluate the causal impact of ranking and agent policy changes under live traffic.
  - Reduced false-positive regressions by ~30%, lowering the risk of misclassifying healthy launches
  - Supported product launches with +2-4% uplift in application starts

- Conceived and shipped a multi-agent job-fit feature on StepStone.de for resume–job semantic alignment and personalised highlight generation.
  - Combined SageMaker-based encoders with OpenAI GPT models via Azure for generation
  - A/B tested impact: 45% reduction in abandonment, +23% completion, +18% qualified applications

- Designed and deployed an agentic feedback triage system, reducing experiment bug report latency from days to minutes.
  - Mitigated indirect prompt injection via a typed schema-based trust boundary with grounding verification at ingestion
  - Reduced ticket-to-resolution time by ~60% and improved experiment attribution accuracy to 95%+

Researcher, Idiap Research Institute **Martigny, Switzerland | Jun 2022 – Dec 2022**

- Built a multilingual NLP pipeline for the EU-funded AI4Media project, processing 500k+ news articles across five languages from 19 European publishers.
  - Combined topic modelling, transformer-based summarisation, and framing classification to analyse narrative structure at scale
  - Identified cross-country media bias patterns across authors, outlets, and nations
  - Supported EU-wide research on framing effects and shifts in reader perception

Data Scientist, EY **Bengaluru, India | Mar 2019 - Sep 2021**

- Developed a hypothesis-driven clustering model on labour economics data (Azure ML) to inform team structuring, targeted hiring, and upskilling strategies across regional business units.
  - Validated via controlled rollout experiments in collaboration with regional business leads, delivering a 1.79% direct cost reduction (~$850k/month)

- Built statistical credit risk models incorporating Monte Carlo simulation to quantify pricing uncertainty and optimise reinsurance and commission structures.
  - Presented scenario analysis to actuarial and finance stakeholders; accelerated premium pricing by +78% and improved profitability by +13.5%; achieved 86.5% BECR with validation via regression and simulation

- Developed a survival analysis system on 1.5M+ real-time sensor datapoints to predict equipment failure and enable proactive maintenance.
  - Partnered with engineering and product teams to translate model outputs into actionable maintenance schedules; reduced remediation time by 95% through early detection and automated alerting; refined thresholds via holdout-style validation

Software Developer, TO THE NEW **Noida, India | Sep 2018 – Mar 2019**

- Built an internal developer-facing self-service portal for cloud provisioning and compute workflows using a REST-based framework
  - Eliminated reliance on external AWS tooling, reducing subscription costs by ~$400k/month