

DEEPAK K. GUPTA

[LinkedIn](#) ◇ [Google Scholar](#)

+971 · 50-155-0628 ◇ guptadeepak2806@gmail.com

Dubai, United Arab Emirates ◇ Resident: UAE (Golden Visa)

AI leader who has built and run the whole stack: the research org, the product, and the company around it. I lead the ML research organisation behind Amazon's internal model compression platform, with scientists across three regions, and I got there by founding a company, taking it from spin-off to acquisition by Amazon in under two years. Before that I ran enterprise AI delivery for an ADNOC joint venture, taking five programmes from feasibility study to production. I set technical strategy, hire and grow scientists, own delivery against hard business targets, and stay close enough to the research to know which bets are real.

LEADERSHIP SNAPSHOT

Org scope	Lead 10+ applied scientists across India, Europe and the US; previously CEO of a venture-backed company and manager of enterprise delivery teams
Strategy	Own the technical roadmap for Amazon's model compression platform, within the hardware–software co-design initiative
Cross-team	Accountable for engagements with product, platform and hardware organisations; translate their constraints into roadmap and drive platform adoption
Business	Founded, funded and exited a company (acqui-hired by Amazon, 2025); PI on research and consultancy grants; enterprise delivery for ADNOC JV
Domains	LLMs and foundation models, efficient AI and edge inference, computer vision, energy and subsurface, healthcare AI

EXPERIENCE

Amazon <i>Applied Science Manager, Model Compression</i>	May 2025 - Present Dubai, UAE
· <i>Scope:</i> Lead the ML research organisation building Amazon's internal model compression platform, a company-wide capability within the hardware–software co-design initiative. Team of 10+ applied scientists across India, Europe and the US. The function was established through the acqui-hire of the company I founded. · <i>Strategy:</i> Own the technical roadmap and research agenda for the platform, and the decision of which compression bets get resourced. · <i>Cross-functional:</i> Responsible for engagements across the organisation – product, platform and hardware teams – translating their deployment constraints and roadmaps into compression requirements, and driving adoption of the platform into their pipelines. · <i>Delivery:</i> Met both organisation-level stretch goals set for compression: a 4× parameter reduction of Qwen2.5-72B at near-parity on public benchmarks, and parameter compression of Qwen3-235B-A22B, a mixture-of-experts model, with faster inference and no measurable quality loss. · <i>Talent and visibility:</i> Grew the team across three regions and positioned it externally, pitching and running a public compression challenge that drew 176 teams. Team research published at CVPR'26 and ICML'26.	

Nyun AI (acqui-hired by Amazon) <i>Co-founder & CEO ← Chief AI Scientist</i>	Aug 2023 - Apr 2025 Bengaluru, India
· <i>Founded and led the company</i> making large foundation models deployable on constrained hardware. Full ownership of strategy, hiring, fundraising, product and customers. Spun out of the research lab I had built at IIT (ISM) Dhanbad.	

- *Business*: Raised **\$158K pre-seed**, booked **\$112K revenue** in the final year, signed clients including US-based deep-tech companies, and secured places in the **Tenity Singapore 2024** and **Hub71 2024** cohorts.
- *Product*: Shipped *Zero*, a commercial model-optimization platform, and *Nyuntam*, an open-source PyTorch compression library, alongside a series of optimized LLMs released publicly.
- *Outcome*: **Acqui-hired by Amazon in May 2025**, with the team and mandate carried into Amazon intact. Research from the company and its originating lab published at **ACL'24**, **IJCAI'24**, **ICLR'24** and **WACV'24**.

AIQ (Presight/ADNOC joint venture)

Jun 2021 - Apr 2025

Data Science Manager ← Lead ← Senior Data Scientist

Abu Dhabi, UAE

- **Promoted twice in four years** within the AI arm of a national energy joint venture.
- *As Manager* (Jun 2023 - Apr 2025): Led a team of four and owned data science delivery across **five enterprise programmes** spanning reservoir portfolio management, corrosion detection and emission forecasting. Set the AI roadmap for each, ran the feasibility studies, and carried delivery through proof-of-concept, pilot and scale-up with domain and business stakeholders.
- *As Senior/Lead* (Jun 2021 - May 2023): Technical lead on an intelligent reservoir portfolio optimization programme; ran the feasibility study, authored the technical proposal, and delivered the PoC and pilot stages.
- *IP*: Programme produced a filed **US patent** on deep-learning-based reservoir portfolio optimization and well placement.

Transmute AI Lab, IIT (ISM) Dhanbad

Jan 2021 - May 2025

Founding Director

Dhanbad, India

- *Built an organisation from zero*: founded the lab within the Texmin Foundation and grew it to **around 20 researchers**, setting its research agenda and five-year vision with the Texmin board.
- *Funding*: Principal investigator on Texmin Foundation research grants and an **Aramco Overseas** consultancy grant applying AI to geophysical problems.
- *Commercial outcome*: Nyun AI was spun out of the lab's research. Lab output published at **CVPR'22**, **ICLR'21**, **ICASSP'23** and **TMLR**.

QUVA Lab, University of Amsterdam

Jan 2019 - Jan 2021

Postdoctoral Researcher

Amsterdam, Netherlands

- Joint **Qualcomm-UvA** research lab. Model efficiency and visual object tracking, with applied projects delivered for TomTom, UMC Utrecht and VUmc. Output at **CVPR'21**, **ICCV'21** and **TPAMI**.

Shell Technology Centre, Royal Dutch Shell

Nov 2017 - Jan 2019

Research Scientist

Bangalore, India

- Seismic Imaging and Analytics. Deep learning to automate seismic processing, and unsupervised detection of energy events in passive microseismic data. Founded *NXplorers*, a machine learning reading group; received a **Shell Special Recognition Award**.

PRODUCTS AND PLATFORMS DELIVERED

- **Amazon model compression platform** - internal production platform serving compression across the organisation.
- **Zero** - commercial platform for automated model optimization (Nyun AI).
- **Nyuntam** - open-source PyTorch compression library: pruning, quantization, distillation, low-rank decomposition.
- **Enterprise AI programmes** - five production programmes for an ADNOC joint venture, feasibility through scale-up.

TECHNICAL CREDIBILITY

- **6 patents** (2 granted, 1 filed, 3 provisional) in compression, binarization, visual tracking and reservoir optimization.
- **60+ peer-reviewed publications** including ICML, CVPR, ICLR, NeurIPS, TPAMI, TMLR and *Nature Scientific Data*.
- Lead organizer, **ICML'26 AdaptFM** workshop on resource-adaptive foundation model inference: 102 accepted papers, 200+ member programme committee drawn from NVIDIA, Qualcomm, Meta, Apple, IBM, MIT and Stanford.
- Lead organizer, **ICCV'23** workshop on Resource-Efficient Deep Learning for Computer Vision. Reviewer for IJCV, NeurIPS, ICML, ICLR and CVPR.
- Associate Professor (20% appointment), UiT The Arctic University of Norway. Supervised 3 PhD, 13 M.Sc. and 20 B.Sc. researchers, a pipeline the lab and company both hired from.

EDUCATION

Delft University of Technology, The Netherlands

2013 - 2017

PhD in Computational Engineering (degree awarded 2019)

Indian Institute of Technology (ISM) Dhanbad, India

2008 - 2013

Integrated B.Sc. and M.Sc. in Geophysics; GPA 8.48/10, **Rank 1, Director's Gold Medal**