Shubham Gupta

Software Engineering Manager - AI · Singapore

(+65) 89408972 | shubhamga2208@gmail.com | goodhamgupta | shubhamgupta2208

Summary

Software Engineering Manager with 9+ years delivering enterprise AI solutions that reduced costs 40% and secured $10M+ in revenue. Proven track record leading high-performance teams across startups and transformation initiatives in finance, healthcare, and government sectors.

Skills

Language Models	Model Training, Model Finetuning & Deployment, VLMs, RAG, PEFT, REFT, vLLM, S-LoRA, Content-Moderation Systems
ML & NLP Systems	PyTorch, JAX, PyMC, Whisper, CLIP, Scikit-Learn, Regression, Random Forests
Data & Infrastructure	SQL, Airflow, ElasticSearch, Kafka, AWS, Docker, Kubernetes, Terraform
Programming	Python, JavaScript, CUDA/C++, Rust, Elixir
Architecture	LLM Serving, Multi-cloud Orchestration, On-device AI, Serverless, Distributed Systems, Event Sourcing

Work Experience

Temus (Temasek backed Digital Transformation Company)

Singapore

Software Engineering Manager - Applied AI

Sep. 2022 - Present

- Led AI implementations across insurance, finance, and government sectors, driving enterprise digital transformation solutions that generated S$12M in annual revenue and winning 7/8 competitive deals through exceptional POC delivery.
- Led AI Investment Analyst Assistant delivery, integrating advanced NLP, knowledge retrieval, and database querying capabilities that reduced due diligence time by 20% and costs by 37%. Achieved 70% accuracy in 4 months (2 months ahead of schedule), driving $10M+ in new projects with 77% client satisfaction.
- Reduced LLM serving costs 40% and POC development time 80% by architecting multi-cloud platform with Kubernetes, vLLM and S-LoRA orchestration. Designed modular infrastructure that became company standard, enabling 3x faster model iterations across 5 product lines.
- Led DialogForge development—a voice-to-voice sales representative training platform using Whisper, Llama3, and Metahuman—reducing onboarding time by 20% for 1,000+ users while improving training outcomes by 30%.
- Advanced implementation of knowledge injection in Llama3 8B models, reducing hallucinations by 95% and improving task accuracy by 50%. Pioneered representation engineering in Mistral/Llama models for production-grade personality control.
- Saved $500K annually by cutting service center call volume 40% through a custom Thai language processing system. Delivered 54% higher intent detection accuracy using fine-tuned Whisper ASR, voice activity detection, and speaker diarization.
- Scaled engineering 8x (3→25 developers) and cut delivery time 40% by implementing agile methods and structured mentorship. Established LLM standards improving code quality 30%, adopted by 4 business units with $1M efficiency savings.

Aaqua (Stealth Social Media Startup)

Singapore

Senior Software Engineer, NLP

Jun. 2021 - Sep. 2022

- Reduced manual content moderation effort by 30% and maintained 85% recall across image perturbations by implementing a CLIP model-based system with robustness to rotation, scaling, cropping and chroma-noise for detecting policy-violating content.
- Improved spam detection recall to 78% from baseline 45% by developing an ML filtering system using Microsoft MiniLM, FastAPI, and Sagemaker trained on open SMS datasets, processing 1M+ messages daily.
- Decreased business report turnaround time by 10% and reduced analytics request backlog by 40% by architecting an ETL platform combining Airflow orchestration, Datahub data management, and dashboards via Superset.

Radicali (Regulatory Compliance Startup)

Singapore

Senior Software Engineer, NLP

Oct. 2019 - May. 2021

- Secured $1M funding by developing a regulation-policy matching system that identified non-compliant clauses and their monetary impact with 65% accuracy, using documents-as-graphs architecture with GPT-2 and knowledge distillation to acheive 20x faster inference.
- Increased web scraping success rates by 45% and reduced infrastructure costs by 25% by implementing a serverless solution using Scrapy with AWS Lambda and Fargate integration.
- Delivered successful product-market fit for 3 major clients representing $500K ARR by managing a 15-person development team, establishing technical delivery processes, and implementing regular stakeholder feedback loops.

Scripbox (FinTech Startup)

Bangalore, India

Senior Software Engineer

Oct. 2016 - Oct. 2019

- Increased company revenue by 10% and achieved 93% prediction accuracy by developing customer lifetime value models using Hierarchical Bayesian methods, enabling targeted retention strategies for highest-value segments.
- Improved customer experience and portfolio transparency for $3M AUM by implementing a portfolio management platform in Elixir that delivered 10x performance improvement, displaying real-time investment returns and metrics.
- Reduced customer onboarding time by 50% and decreased document processing errors by 35% by architecting an automated pipeline combining YOLOv2 object detection with Tesseract OCR for identity and financial document verification.

Zopper (Ecommerce Startup) Bangalore, India

Software Engineer Jan. 2016 - Oct. 2016

- Improved order processing efficiency by 13% and reduced system downtime by 40% by implementing a LinkedIn Databus-based database migration system that enabled zero-downtime schema changes across the e-commerce platform.
- Increased retail sales by 20% and expanded product catalog by 35,000 SKUs by developing a catalog extension platform that enabled national distributor product sourcing and real-time inventory synchronization.

Selected Projects

Open-Source Contributor Singapore

GitHub, Elixir Packages Aug. 2016 - Present

- Open source contributor to Nx, OpenMined and Snowplow Analytics. Packages have over **250k** downloads

ExStan: Elixir <> Stan for Probabilistic Modeling Lisbon, Portugal

bit.ly/exstan-elixirconf Apr. 2024

- Presented a talk on Probabilistic Programming using ExStan, Stan SDK in Elixir at ElixirConfEU 2024

AI Driven Patient Engagement in Healthcare Singapore

bit.ly/on-device-llm-healthcare Jan. 2024

- Improved healthcare access for 100+ patients in low-resource regions by engineering a mobile app with optimized on-device LLM (3B params) using token logit confidence scoring and cloud fallback, reducing latency 65% while maintaining 90% accuracy.

State Space Models for Language Modeling Singapore

bit.ly/mlsg-ssm Feb. 2024

- Presented State Space Models (SSMs), an alternative architecture for Language Modelling, in the ML Singapore Meetup

Hierarchical Bayesian CLV Model Moscow, Russia

bit.ly/hierachical-clv Aug. 2019

- Developed and presented at DeepBayes2019, a novel Bayesian hierarchical model for CLV prediction that integrates customer demographics, enabling data-driven marketing strategies.

Accomplishments

Anthropic Bsides Security Challenge Online

bit.ly/anthropic-bsides May. 2024

- Earned $500 in API credits and recognition from Anthropic's security team by successfully solving their BSides challenge through advanced steganography detection and LLM jailbreak mitigation techniques.

End-to-End Attention based Image Captioning Singapore

bit.ly/attention Apr. 2021

- Implemented as part of a course project at NUS, and finished in top 10% of the leaderboard in the Kaggle contest.

Education

NUS(National University of Singapore) Singapore

Masters in Computer Science, Artificial Intelligence Aug. 2020 - Dec. 2022

CGPA: 4.0/5.0

Amrita School of Engineering Bangalore, India

Bachelor of Technology in Computer Science Aug. 2012 - May. 2016

CGPA: 7.96/10.0

Online Courses

Mathematics for Machine Learning	MIT
Mastering LLMs for Developers & Data Scientists	Maven
Fast AI	fast.ai
Machine Learning	Udacity
Computer Vision	Stanford
Deep Learning for NLP	Stanford
Distributed Systems	MIT, Empowered Coder