Shubham **Gupta**

Software Engineer · ML Engineer

shubhamga2208@gmail.com | shubhamg.in | goodhamgupta | shubhamgupta2208

Summary

Results-driven Software and ML Engineer with 8 Years of experience in effective AI products. Proven track record of building and delivering solutions to create business value. Passionate about open-source and on-device local AI.

Skills

NLP	Language Modeling, Model Finetuning & Deployment, PEFT, REFT, Content-Moderation Systems
Machine Learning	Regression, SVM, K-Means, Random Forests, TF-IDF
Frameworks	PyTorch, Jax, PyMC
DevOps	AWS, Docker, Kubernetes, Terraform
Data Engineering	SQL, Airflow, ElasticSearch, Kafka
Programming	Python, JS, Java, Rust, Elixir

Work Experience

Temus — *Singapore*

Software Engineering Manager, Data and AI — *Oct. 2022 - Present*

- Reduced financial due diligence time by 20% and costs by 37% by leading development of an AI financial assistant that combined LLMs with RAG and SQL AST parsing, achieving 70% accuracy through on-premise deployment of the SQLCoder Text-to-SQL model.
- Improved Thai language intent detection accuracy by 54% by architecting a multilingual ASR system utilizing fine-tuned Whisper, Voice Activity Detection, and Speaker Diarization with Llama-based transcript correction.
- Scaled engineering team from 3 to 25 developers across 3 squads by implementing agile methodologies, mentoring programs, and technical development paths, resulting in 40% faster project delivery and 30% improvement in code quality metrics.
- Reduced sales representative onboarding time by 20% across 1,000+ users by leading development of DialogForge, a real-time voice-to-voice training platform combining Whisper, Llama3, Elevenlabs, and Unreal Engine Metahuman.
- Decreased LLM serving costs by 40% by architecting LinguaForge, a multi-cloud platform utilizing Kubernetes, SkyPilot, and vLLM with S-LoRA orchestration for optimized adapter management.

Aaqua — *Singapore*

Applied Research Engineer, NLP — *Jun. 2021 - Sep. 2022*

- Reduced manual content moderation effort by 30% by implementing a CLIP model-based content moderation system that automatically detects policy-violating content across the platform.
- Improved spam detection recall to 78% by developing an ML-based spam filtering system using Microsoft MiniLM model, FastAPI, and Sagemaker, trained on open SMS spam datasets.
- Decreased business report turnaround time by 10% by architecting an ETL platform that combines Airflow orchestration, Datahub data management, and Superset visualizations.

RADICALi — *Singapore*

Senior Data Scientist — *Oct. 2019 - May. 2021*

- Improved document analysis accuracy to 65% and achieved 20x faster inference by leading development of an platform combining GPT-2 and documents-as-graphs architecture with a 0.45 silhouette score.
- Increased web scraping success rates by 45% and reduced costs by 25% by architecting a serverless solution using Scrapy with AWS Lambda and Fargate integration.
- Delivered successful product-market fit for 3 major clients by managing a 15-person development team, establishing technical delivery processes, and implementing regular stakeholder feedback loops.

Scripbox — *Bangalore, India*

Product Engineer and Data Scientist — *Oct. 2016 - Oct. 2019*

- Increased company revenue by 10% and achieved 93% prediction accuracy by developing customer lifetime value models using Hierarchical Bayesian methods and Prophet forecasting.
- Improved customer onboarding speed by 50% by architecting an automated document processing pipeline combining YOLOv2 with Tesseract-based OCR.
- Accelerated data analytics capabilities to real-time processing by implementing an event sourcing architecture using Snowplow, Kinesis, and Redshift integration.

Zopper — *Bangalore, India*

Software Engineer — *Jan. 2016 - Oct. 2016*

- Improved order processing efficiency by 13% and increased retail sales by 20% by implementing LinkedIn Databus based database migration system and a catalog extension platform that enabled national distributor product sourcing.

Selected Projects

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

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

AI Driven Patient Engagement in Healthcare
Singapore

BIT.LY/ON-DEVICE-LLM-HEALTHCARE
Jan. 2024

- Engineered mobile app with optimized on-device LLM (3B params) and confidence-based cloud fallback, using token logits, to large model running on the cloud
- Successfully piloted with 100 users, enabling accessible medical guidance in low-resource regions

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.

Hierarchical Bayesian CLV Model
Moscow, Russia

HTTPS://BIT.LY/HIERACHICAL-CLV
Aug. 2019

- Developed and presented at DeepBayes2019, a novel Bayesian hierarchical model for CLV prediction that integrates customer demographics, achieving 37% improvement in valuation accuracy and enabling data-driven marketing strategies

Open-Source Contributor
Singapore

GitHub, Elixir Packages
Aug. 2016 - Present

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

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*