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*

- Led development team in building an AI financial assistant that reduced due diligence time by 20% and cut costs by 37%. System combined LLMs with RAG and SQL AST parsing for structured and unstructured data analysis, achieving 70% accuracy through on-premise LLM deployment.
- Architected multilingual ASR system for Thai language processing using fine-tuned Whisper, Voice Activity Detection, and Speaker Diarization, boosting intent detection accuracy by 54%. Enhanced robustness through Llama-based transcript correction pipeline.
- Architected and led development of a large-scale logistics optimization platform that orchestrates route planning for 3,000+ trucks, leveraging Optaplanner scheduling problems under domain constraints, resulting in a 50% efficiency improvement in fleet utilization & delivery times.
- Led development team in building DialogForge, a real-time voice-to-voice training platform using Whisper, Llama3, and Elevenlabs, integrated with Unreal Engine Metahuman for realistic avatar interactions. Deployed to 1,000+ sales representatives, reducing onboarding time by 20%
- Architected LinguaForge, a multi-cloud LLM platform using Kubernetes, SkyPilot, and vLLM integration, reducing serving costs by 40%. Implemented S-LoRA orchestration for efficient adapter management, optimizing model performance across deployments.

Aaqua — *Singapore*

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

- Implemented a content-moderation system based on the OpenAI CLIP model to detect policy violating content on the platform, thereby decreasing manual moderation effort by 30%.
- Implemented a system to detect spam messages, based on the Microsoft MiniLM model and an open SMS spam dataset, using FastAPI and Sagemaker, increasing recall to 78%.
- Implemented a platform to perform ETL of data from multiple sources, using Airflow for orchestration and Datahub for data management, and Superset for visualisations, reducing TAT for business reports by 10%.

RADICALi — *Singapore*

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

- Lead development of an advanced document and topic identification systems using BERT, Query Re-ranking, and documents-as-graphs, achieving a silhouette score of 0.45 and 65% accuracy in summarization, with a 20x increase in inference speed through model optimization.
- Implemented a serverless scraping architecture using Scrapy, AWS Lambda, and Fargate, increasing scraping success rates by 45% and reducing costs by 25%.
- Managed a 15-person development organization, overseeing technical delivery while cultivating client partnerships and steering product direction to ensure market fit

Scripbox — *Bangalore, India*

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

- Developed forecasting and customer segmentation models, achieving 93% accuracy in revenue and customer predictions using Prophet and increasing revenue by 10% through a Hierarchical Bayesian CLV model.
- Enhanced data processing and analytics capabilities by implementing real-time event sourcing with Snowplow, Kinesis, and Redshift, and improving customer onboarding by 50% with a YOLO DL model and OCR-based document processing pipeline.

Zopper — *Bangalore, India*

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

- Implemented a software database AESOP(LinkedIn) which helped process changes from one type of data store to another reliably. Ingested order processing efficiency by 13%.
- Implemented 'Catalog Extension', a platform to allow retailers to source products from national distributors thereby increasing overall sales by 20%.

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*