# Santhosh Muthukrishnan *Software Engineer*

ie.santhosh.m@gmail.com Bengaluru, India linkedin.com/in/thataintme

github.com/thataintme

## Profile

AI-focused Software Engineer building production LLM, Retrieval-Augmented Generation (RAG) and multimodal applications. Experience delivering serverless cloud architectures, fine-tuning LLMs (QLoRA, GRPO), and translating stakeholder requirements into deployed AI solutions.

## Education

**Master of Science (MSc) in Data Science and Analytics** 08/2024 – 10/2025 | Maynooth, Ireland
*Maynooth University*
First-class honours 1:1, 81.8%

**Bachelor of Technology (BTech) in Computer Science Engineering** 06/2018 – 06/2022 | Vellore, India
*Vellore Institute of Technology*
First class honours 1:1 87.5%

## Professional Experience

**AI Research Assistant** 05/2026 – 08/2026 | Maynooth, Ireland
*Maynooth University*

- Benchmarked 8 transformer and 4 vision-language models across 3 datasets for prostate cancer Gleason grading. The best model reached **82.56%** patch-level accuracy.
- Built ingestion and patching pipelines that make gigapixel Whole Slide Images (multi-GB each) trainable on a GPU **HPC** cluster, using **Slurm**.
- Turned the research pipeline into a diagnostic web app where users upload WSIs or patches and get predictions from any of the trained ViT and VLM models. Frontend hosted on Cloudflare; models served via Hugging Face Spaces.
- Built a reproducible experiment framework (data and library versioning, configs, logging) and tuned Slurm jobs to cut queue time and wasted compute.

**Research Intern** 05/2025 – 12/2025 | Dublin, Ireland
*ADAPT Centre*

- Built multimodal pipelines combining clinical tabular data (demographics, biomarkers) with 3D MRI imaging, for an LLM-based diagnosis system.
- Cleaned sparse longitudinal clinical records with missing values, misaligned timestamps and schema mismatches.
- Built an end-to-end XGBoost pipeline (features, training, validation) that achieved **$R^2 = 0.664$**.
- Used vision-embedded large language models to generate structured text annotations at scale for downstream training.

**Software Engineer** 01/2022 – 06/2024 | Chennai, India
*Codemantra*

- Owned backend microservices for an AI-powered PDF accessibility product, from design through deployment. Stack: Spring Boot, Flask, AWS Lambda/SAM, SNS/SQS.
- Made PDF-processing Lambdas up to **12x faster**, which **cut timeouts by 89%** and **cut compute costs by 60%**. Added local fallbacks for jobs that exceed Lambda limits.
- Sped up local server processes by an average of **50%+**, reducing peak PDF processing time on local servers from 45 to 22 minutes.
- Built annotation tooling that produced labeled image-JSON training data, plus the pipeline that serves rendered pages to the models for inference.
- Built a complete PDF accessibility verifier based on industry accessibility standards.
- Automated internal workflows, which **raised ticket closure by 30%**.

**Java Developer** 09/2021 – 12/2021 | Bengaluru, India
*Dockabl*
- Built REST APIs and resolved bug tickets for an HR analytics platform used by managers and HR teams.
- Optimised backend code, wrote unit tests, and introduced Java Checkstyle rules to improve code quality.

## Projects

**Biomedical Literature RAG System**
*MedCPT, FAISS, BM25, Hugging Face, Torch, Gemini Flash*
- Built a **Retrieval Augmented Generation** QA system over PubMed abstracts that fuses **BM25** and **MedCPT** dense retrieval with **Reciprocal Rank Fusion** and a **cross-encoder reranker**. **Gemini Flash** generates answers that cite the retrieved abstracts by PMID.
- Evaluated on **5,729 BioASQ questions**. Hybrid retrieval raised **Recall@10 from 0.655 to 0.683** and **MRR@10 from 0.859 to 0.869** over a BM25 baseline, and reranking pushed Recall@10 to **0.699**.
- Deployed on **Hugging Face Spaces (ZeroGPU)** with **end-to-end p50/p95 latency of 2.3 s / 4.7 s**. Retrieval runs in **21 ms** and reranking 100 candidates in **273 ms** at p95; LLM generation contributes to about 90% of latency.
- Halved the reranker's input per candidate by truncating to fewer (256) tokens, at about 1% relative recall loss.

*Live demo at bioasq.santhosh.world*
*(static frontend on Cloudflare, retrieval and generation on Hugging Face ZeroGPU backend).*

**LLM Code Review and Repair (NVIDIA OpenCodeReasoning)**
- Fine-tuned Qwen2.5-Coder-7B to detect bugs, explain them and generate fixes. Trained on NVIDIA's OpenCodeReasoning dataset of generated solutions and reviewer critiques.
- Ran two-stage training: QLoRA supervised fine-tuning using Unsloth, with GRPO (group size 4) with an **execution-based reward** against test cases.

**Cloud-Native PDF Asset Extraction System**
*AWS Lambda, S3, DynamoDB*
- Built a serverless AWS pipeline to extract image-caption pairs and metadata from PDFs at scale.
- Uses Lambda for processing, S3 for assets, and DynamoDB for metadata, with decoupled services and automated storage lifecycle management.

**YTBUDDY: Automated YouTube Shorts Pipeline**
*Google Cloud App Engine, Google Drive API, YouTube Data API, OAuth 2.0, [Python/Flask]*
- Built a serverless web app on Google App Engine that pulls long-form videos (streams, vlogs) from Google Drive, segments them into 30–90 second clips, and uploads them to YouTube as Shorts.
- Uses Google Drive API and YouTube Data API.
- Built a dashboard with OAuth2 for choosing videos, approving clips, and scheduling uploads.