

Shubh Sareen

shubhsareen@iitb.ac.in | [Shubh Sareen | LinkedIn](#) | [Shubh Sareen | Github](#)

Education

IIT Bombay – B.Tech in Chemical Engineering with Minor in Machine Intelligence and Data Science (CPI: 8.72) Jul 2024 – Jul 2028

Experience

Research and Engineering Intern, Autonise – Bengaluru, Karnataka May 2026 - Jun 2026

• SVG Generation Infrastructure:

- Replaced direct **LLM-generated SVG code** with an **LLM → Python + Library → SVG** architecture, eliminating geometry hallucinations and enabling reliable SVG generation and reducing SVG-generation costs by **95%**
- Benchmarked **Qwen3.6-35B** against **Claude Sonnet 4.6** for structured SVG generation, enabling migration to local inference while preserving generation quality and consistency

• Evaluation & Retrieval Research:

- Developed and benchmarked LLM-based question-difficulty ranking pipelines using Round-Robin, Merge-Sort, and Swiss-Tournament evaluation frameworks, identifying limitations that informed ranking-system design
- Designed a **similar-question** discovery pipeline that restricted **pairwise comparisons** to difficulty-ranked local neighborhoods, enabling scalable similarity-link generation and efficient retrieval of related educational content
- Integrated similar-question **recommendations** into production **doubt-solving** and test-analysis interfaces, **improving** content exploration and practice continuity for students on autonise.com/doubts

• Document Understanding & Extraction:

- Deployed an **end-to-end** document digitization pipeline (**Django, Angular, MongoDB**) with human feedback
- Generated a **33K-image** dataset (varied fonts/layouts/diagrams) with **LLM filtering** and bounding-box refinement
- Re-engineered a vision model for **figure-extraction**, fixing a 4× val-loss anomaly to **96.1% mAP50 / 88.9% mAP50-95**

Research Intern, Info Me – Remote

Jan 2026

- Architected a two-stage pipeline combining embedding retrieval and **LLM reranking** for sparse user data
- Built user vectors from category-conditioned embeddings, retrieved **top-k** candidates via **cosine** similarity
- Identified breakdown in category-overlap metric due to coarse embedding clusters, redesigned architecture
- Split LLM call into **summarizer** and **reranker agents** on locally hosted **Llama 3B**, reducing context and hallucination

Key Projects

YouTube Playlist Summarizer

<https://yt-helper.shubhsareen.com>

- Architected and deployed a **multi-LLM** YouTube learning assistant on **Azure** Container Apps, transforming videos and playlists into searchable repositories with **RAG-based** Q&A, study-note synthesis, flashcard generation, and quizzes
- Designed a **semantic retrieval** pipeline using transcript chunking, **cosine-similarity** search, and **LRU** embedding caches, reducing redundant embedding generation and improving query latency for repeated queries
- Engineered a **provider-agnostic** orchestration and ingestion layer with OpenRouter/Gemini failover, **playlist-scale** transcript processing, per-video fault isolation, and resilient handling of API failures, outages, and rate limits

Vision-to-Text

[ShubhSarin/Vision-to-text-SOC](#)

- Developed **PyTorch CNNs** for image **classification** and applied transfer learning with **ResNet50** on CIFAR-10
- Trained a 2-layer **LSTM** language model in PyTorch on **500K+** tokens with GloVe-100D embeddings for text generation
- Architected **vision-to-text** model with ResNet-50 **encoder**, attention-based LSTM **decoder**, and GloVe-100D

Skills

Programming Languages: Python, C++, JavaScript

AI & ML: Deep Learning, NLP, LLM, Supervised and Unsupervised Learning

Frameworks: PyTorch, LangChain, LangGraph, React, Django, Angular