Yashmit Vaish

yvaish_be24@thapar.edu | +91 93359 64110 | linkedin.com/in/yashmit-vaish | github.com/YashmitVaish

Education

Thapar Institute of Engineering and Technology, Patiala, India 2024 – 2028 (Expected)
Bachelor of Engineering in Computer Engineering CGPA: 9.28/10

Experience

Clyra Remote | Dec 2025 – Present
Generative AI Intern - AI-powered content platform

- Owned the AI system architecture as sole engineer - designed and shipped production RAG pipelines (LangGraph + Elasticsearch + Neo4j) with async job processing and retry/error handling, load-tested with Locust to sustain **30–40 content files/day**.
- Slashed LLM inference costs by **30%** and reduced redundant re-processing by **70–80%** by implementing model routing and response caching - routing simpler tasks to cheaper models and caching repeated retrieval calls.
- Improved RAG output quality by **30% (ROUGE)** by redesigning the chunking strategy, tuning the retrieval layer across Elasticsearch and Neo4j, and optimizing LangGraph agent orchestration then validated against a held-out content benchmark.

Projects

Campus Permission Portal Thapar Institute of Engineering and Technology
FastAPI, React, PostgreSQL, Redis, Supabase, Nginx, SMTP, AWS *2-person team | Production*

- Gathered requirements from campus faculty/administration to scope multi-role approval workflows, then built and deployed a full-stack permission platform (FastAPI, React, PostgreSQL) that scaled to **5,000+ peak concurrent users** (observed peak, usage logs) with **zero downtime**.
- Architected an async FastAPI + Redis job-queue backend, fronted by Nginx as reverse proxy, to absorb request bursts during the portal's daily 4–5 hour active-use window, sustaining **approx. 3,000 average daily page views**.
- Reduced status-inquiry follow-ups to faculty from **50+ to 1–2 per week** by integrating a **REST API**–driven transactional email service with direct status links and real-time visibility.
- Wrote unit tests covering core approval-workflow logic and deployed the platform on **AWS**, sustaining reliable operation across peak admission-cycle traffic.

Pdf2Video github.com/YashmitVaish/Pdf2Video
Manim, LangChain, Ollama, vLLM, Piper, CrewAI

- Engineered a multi-agent pipeline that generated **100+** narrated, animated video lectures from academic PDFs using Manim and fully offline LLM inference - removing GPU/cloud dependency and cost.
- Shortened per-video generation time by **15%** by implementing multi-process task scheduling and CPU-only TTS via Piper.
- Orchestrated **6** specialized CrewAI agents (orchestrator, extractor, script generator, scene generator, audio renderer, final-pass compositor) across a **microservices** task architecture, using LangChain shared context for end-to-end automated production.

Media Processing & Normalization Pipeline github.com/YashmitVaish/vdo-process
Python, FastAPI, Redis, MinIO, FFmpeg, MongoDB, MediaMTX, Docker

- Developed an async video-processing microservice achieving **approx. 6-second** end-to-end latency for live-stream processing, using Redis job queues, MinIO object storage, and FFmpeg-based normalization; containerized via Docker Compose for one-command deployment.
- Implemented signed upload URLs, FFprobe metadata analysis, multi-pass normalization (resolution, FPS, audio loudness), and crossfade merge support within a **CI/CD**-driven pipeline for a reliable, horizontally scalable ingest service.

Leadership

ACM Student Chapter, Thapar Institute of Engineering and Technology Patiala, India
General Secretary Aug 2026 – Present

- Led a structured team of **100+** members across coordinated sub-teams to plan and execute **4** chapter-wide events, including a freshman orientation, a hackathon, and multiple technical workshops.
- Owned end-to-end event operations - planning, logistics, team coordination, and on-ground execution - ensuring smooth, on-schedule delivery across all events for the student community.

Technical Skills

Languages: Python, Go, C++, C
AI/ML & LLM Engineering: RAG, LangChain, LangGraph, CrewAI, vLLM, Ollama, llama.cpp
Backend & Infrastructure: FastAPI, Django, Gin, REST API, Microservices, Redis, Celery, Docker, Nginx, Cloudflare Tunnels, CI/CD, Async/Concurrent Systems
Cloud: AWS | **Databases & Vector Stores:** PostgreSQL, MySQL, MongoDB, Pinecone, Neo4j, Elasticsearch, Supabase
Tools & Platforms: Git, GitHub, Locust, Vercel, VS Code